INTEGRATION OF NMR AND SAXS WITH ATOMISTIC SIMULATIONS FOR CHARACTERIZING THE STRUCTURE AND DYNAMICS OF MULTI-DOMAIN PROTEINS

by

Karl Thomas Debiec

B.S. in Chemistry and Molecular Biology, University of Pittsburgh, 2009

M.Ch.E. in Chemical Engineering, Carnegie Mellon University, 2010

Submitted to the Graduate Faculty of

The Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Molecular Biophysics and Structural Biology

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS & SCIENCES

This dissertation was presented

by

Karl Thomas Debiec

It was defended on

July 21st, 2017

and approved by

Angela M. Gronenborn, Ph. D., Distinguished Professor of Structural Biology

Lillian T. Chong, Ph. D., Associate Professor of Chemistry

W. Seth Horne, Ph. D., Associate Professor of Chemistry

David A. Case, Ph.D., Distinguished Professor of Chemistry and Chemical Biology, Rutgers University

Dissertation Advisor: Angela M. Gronenborn, Ph. D., Distinguished Professor of Structural Biology

Dissertation Advisor: Lillian T. Chong, Ph. D., Associate Professor of Chemistry

Integration of NMR and SAXS with Atomistic Simulations

for Characterizing the Structure and Dynamics of Multi-Domain Proteins

Karl Thomas Debiec, Ph. D.

University of Pittsburgh, 2017

In the seven decades since the first atomic-level structures of biomolecules were determined, the development and application of novel research methods has led to an advanced understanding of biological functions at the molecular level. In addition to experimental methods, key advances have been spurred by computer simulations, which provide an *in silico* representation of accumulated prior knowledge of biomolecular structure and dynamics. These models can be used both (i) as a complement to experimental results, filling in the gaps where experimental information is not accessible, and (ii) as complete representations, directing future research. Critically, the validity of either application depends on the accuracy of the models used. In this work, I aspired to combine computational and experimental methods to characterize the structure and dynamics of the flexibly linked two-domain protein MoCVNH3. In Chapter 1 I describe my motivation, and the suspected simulation artifacts observed in our preliminary simulations, which led me to investigate how accurately simulation models represent salt bridge interactions. Chapter 2 details my comparison of current models ("force fields"), for which significant variation but consistent overstabilization of salt bridges was discovered. This work motivated the development of a new force field, AMBER ff15ipq, which corrects, to some degree, the overstabilization and introduces extensive improvements, described in Chapter 3. Finally, in Chapter 4, I applied this new force field in simulations of MoCVNH3, for which I collected extensive experimental data leading to the determination of a structural ensemble. I validated the simulations against the

experimental data set, and identified further directions for improvement. Overall, the work presented here demonstrates the power of integrating experimental and computational methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 MOLECULAR BIOPHYSICS OF FLEXIBLY LINKED MULTI-DOMAIN PROTEINS

1.1 INTRODUCTION

Multi-domain proteins in which the connected domains each fold and function independently are prevalent in nature.[1,2] Such proteins, through spatial and temporal coordination of their varied functional units, are capable of executing specific and tailored activities in catalysis, signaling, regulation of gene expression, and other cellular processes.[3] The individual domains are connected by inter-domain linkers whose length and composition enable them to adopt orientations that have evolved for specific biological activities and functions.[4,5] In many cases, the linkers are highly flexible, allowing the domains to adopt numerous inter-domain orientations, from which the selection of functional competent conformations may occur.[6] While most multi-domain proteins are linked linearly in sequence, roughly one tenth possess domain insertions where a 'guest' domain is implanted into a loop of a 'host' domain, such that the two domains are connected by a pair of inter-domain linkers.[7]

Characterization of the relative domain orientations within multi-domain proteins has been challenging by traditional structural biology techniques, such as X-ray crystallography, due to the inherent flexibility of inter-domain linkers, lack of density for certain segments of the polypeptide chain, and influence of crystal packing on the positioning of domains. On the other hand, multi-

domain proteins represent intriguing targets for integrative structural biology approaches, which combine results from experiments and computer simulations[5,8] by either (i) computationally generating a large ensemble of potential structural models and subsequently filtering the models based on agreement between computational and experimental data, or (ii) explicitly biasing the generation of structural models in accord with the experimental data. Such approaches have been particularly useful for studying flexibly linked multi-domain proteins and protein complexes,[9–15] often integrating data from NMR, SAXS, X-ray crystallography, and other experimental techniques into a single structural model. Critically, the validity of any approaches aimed at bridging the gaps between experimentally accessible and computationally generated data depends on the accuracy of the biomolecular force fields used in the computations, which dictate sampling of the conformational space for the entire system.

1.2 THE MOCVNH3 PROTEIN

An ideal test system among flexibly linked multi-domain proteins is the relatively small, two-domain protein MoCVNH3 that has been structurally characterized by our group using both NMR spectroscopy and X-ray crystallography.[16,17] MoCVNH3 is a domain-insertion protein in which a 'guest' LysM domain is inserted into a surface loop of a 'host' Cyanovirin-N Homology (CVNH) domain, positioning the LysM domain between the two pseudo-symmetric halves of two-lobed CVNH domain (Figure 1.1).[18] This protein is found in *Magnaporte oryzae*, an ascomycete fungus that causes rice blast disease, the most devastating infection of cultivated rice, which destroys crops in unprecedented amounts worldwide.[19] Functionally, both CVNH and LysM are carbohydrate-binding domains; CVNH binds to mannose sugars, while LysM interacts with

GlcNAc-containing carbohydrates such as peptidoglycan and chitin.[20,21] The binding of carbohydrates by each domain in MoCVNH3 is independent of the other, with no communication between the domains.[22] While the wild-type protein could not be crystallized, complete removal of the inter-domain linkers yielded a construct that crystallized and maintained the ability to bind both carbohydrate ligands. A comparison of the resulting crystal structure with the NMR structure of wild-type MoCVNH3 revealed that the absence of the linkers has no effect on the structures of the individual domains.[17] However, although the domain structures of wild-type MoCVNH3 were solved to high resolution by NMR, no fixed relative domain orientations were compatible with the solution data, due to the lack of inter-domain restraints.[22]

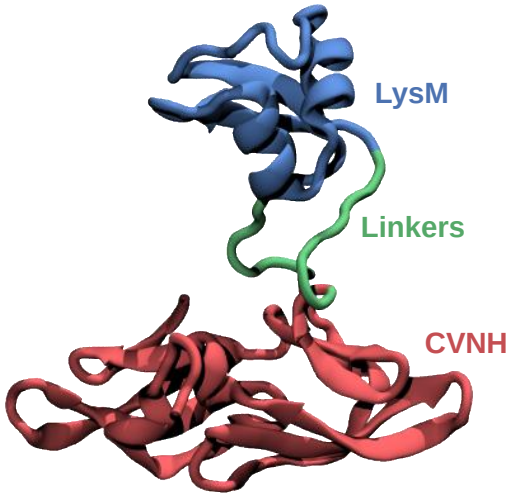

Figure 1.1. Lowest energy structure of the solution NMR ensemble of the two-domain protein MoCVNH3.[22] The CVNH domain is shown in red, the LysM domain in blue, and the inter-domain linkers in green. Other structures within the ensemble have different inter-domain orientations.

1.3 SIMULATION AND SUSPECTED ARTIFACTS

Since the details of the inter-domain orientations sampled by MoCVNH3 were inaccessible to our NMR data, we instead turned to all-atom molecular dynamics (MD) simulation in explicit solvent. MD simulation can, in principle, provide a view of biophysical processes with atomic spatial resolution and femtosecond temporal resolution. However, the accuracy of MD simulation is limited by (i) the simulation duration, restricting the accessible biological processes and the quantitative precision of the results, and (ii) the force field, restricting the results' accuracy and the qualitative confidence that may be placed in them. Over the last two decades improvements and hardware and algorithms have increased the maximum achievable duration of MD simulations from one microsecond[23] to over one millisecond.[24] A key advance was the development of Anton, a supercomputer designed solely and specifically for running MD simulations, whose optimizations enable it to simulate roughly 3 orders of magnitude faster than the conventional resources available at the time of its development.[25] We were granted an allocation on Anton to simulate MoCVNH3, making it feasible to seek the multi-microsecond duration we anticipated would be needed to characterize the distribution of inter-domain orientations.

Increases in accessible simulation duration drive advancements in the accuracy of force fields, as longer simulations reveal artifacts not immediately apparent in shorter ones. In order to optimally make use of our Anton allocation, we used conventional resources to run a series of 1-μs tests using several force fields and water models, eventually selecting the AMBER ff99SB-ILDN force field and TIP4P-Ew water model for our production simulation.[26,27] At the time, AMBER ff99SB-ILDN represented the latest generation of the AMBER force field lineage, including improvements to backbone and side-chain torsion parameters,[26,28] while TIP4P-Ew offered accurate reproduction of the properties of water alongside improved reproduction of NMR

observables.[27,29] Based on our NMR results,[22] our expectation was that the two domains would sample a variety of inter-domain orientations over the course of our simulation. In our 1-μs test, this was what we observed.

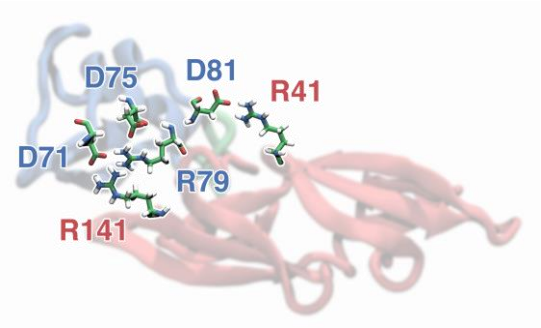

Figure 1.2. Fixed inter-domain orientation of MoCVNH3 observed in a 1-μs MD simulation of MoCVNH3 using the AMBER ff99SB-ILDN force field and TIP4P-Ew water model. Residues involved in salt bridge interactions present at the inter-domain interface are labeled and shown in licorice.

However, when we ran our simulation of MoCVNH3 on Anton, we found that after ~30 ns the two domains of MoCVNH3 adopted a conformation in which the LysM domain latched onto the CVNH domain, and remained in this orientation for the remainder of the 1-μs simulation. This extensive sampling of a stable, fixed inter-domain orientation contrasted with our expectations based on our NMR data. We examined the inter-domain interactions between the two domains, finding that a series of salt bridges, or pairs of amino acids whose oppositely charged side-chains are within hydrogen-bonding distance,[30] were a key part of the complex (Figure 1.2). Subsequent reexamination of our test simulation showed that similar contacts were made between the two domains at the end of that simulation. The two domains did not come into close contact for the first 850 ns of the simulation, but once they did, similar salt bridge interactions were formed for the remaining 150 ns. While our 1-μs simulations were insufficient to conclude that the simulation model contradicts experiment, it led us to suspect that the result might be an artifact of the chosen

force field and water model. We therefore set out to test the accuracy of salt bridge interactions in AMBER ff99SB-ILDN/TIP4P-Ew and other force field/water model combinations, as described in the next chapter.

2.0 THE ACCURACY OF SALT BRIDGE INTERACTIONS IN BIOMOLECULAR FORCE FIELDS

This chapter is based on a research article previously published as: Debiec, K. T.; Gronenborn, A. M.; Chong L. T. Evaluating the strength of salt bridges: A comparison of current biomolecular force fields. *Journal of Physical Chemistry B*, 2014, 118, 6561-6569.

2.1 INTRODUCTION

Recent advances in computer hardware and software have greatly extended the time scales that can be covered by biomolecular simulations. These longer time scales (beyond nanoseconds) are essential for the rigorous evaluation of current biomolecular force fields. One important characteristic of these force fields is the ability to accurately model the formation of salt bridges, or pairs of amino acids whose oppositely charged side-chains are within hydrogen-bonding distance in proteins.[30] However, it has long been suspected that the forces between oppositely charged amino acids are overly attractive in molecular dynamics (MD) simulations with current biomolecular force fields, and there have been a number of efforts to reduce this artifact in the improvement of various force fields.[31–33] Previous theoretical studies have analyzed the contribution of salt bridges to protein or protein-protein complex stability using both implicit[34–38] and explicit modeling of solvation.[39,40] Others have studied salt bridges using amino acid analogues,[41–49] often using biasing techniques in the simulations.[42–44] More recently, a comprehensive comparison of force field/water model combinations was conducted for salt bridge

interactions between the amino and carboxyl groups of zwitterionic amino acids using extensive simulations in explicit solvent on the microsecond (μs) time scale.[50]

Here, we evaluated six biomolecular force fields for their ability to accurately model the strengths of salt bridges between the side chains of oppositely charged amino acids by unbiased, microsecond-scale MD simulations in explicit solvent. In particular, we directly compared current AMBER, CHARMM, and OPLS force fields in simulations of association between the side-chain analogues of three different pairs of amino acids, Arg/Asp, Lys/Asp, and His(+)/Asp. We further tested one of the pairs, Arg/Asp, by simulating association of blocked amino acid dipeptides. In addition, we evaluated the influence of the solvent model on the strengths of the salt bridges by simulating the side-chain analogue pairs using a selection of different force field/water model combinations. To our knowledge, our microsecond-scale simulations provide the most extensive sampling of salt bridge formation to date, yielding thousands of association/dissociation events, permitting quantitative comparisons, both between the force fields and with experiment. Our results reveal considerable variability among the current force fields in terms of the resulting strengths of salt bridge interactions as well as differences from experimental data.

2.2 METHODS

2.2.1 Preparation of starting models

We modeled the formation of salt bridges between the following pairs of oppositely charged amino acids using side-chain analogues: Arg/Asp (guanidinium cation/acetate anion), Lys/Asp (butylammonium cation/acetate anion), and His(+)/Asp (imidazolium cation/acetate anion). Our

systems were constructed to be consistent with the experimental conditions under which the equilibrium association constants (K_A) of guanidinium acetate and butylammonium acetate have been measured,[51] i.e. using the same concentrations (0.9 M guanidinium and 0.02 M acetate, which corresponds to 100 molecules of guanidinium and two molecules of acetate in the presence of ~18,000 explicit water molecules). To ensure a net charge of zero, we included 98 chloride ions (the same counterion that is present in the experiments). The same concentrations of the cation, anion, and chloride ions were also used for the model systems consisting of butylammonium/acetate and imidazolium/acetate. Starting models for these simulations were constructed using the Packmol software package,[52] immersing the appropriate number of side-chain analogues in periodic, cubic boxes of explicit solvent. For the Arg/Asp salt bridge, we also used blocked amino acid dipeptides (acetyl—arginine—N-methylamide and acetyl—aspartate—N-methylamide) to model salt bridge formation. Only a single copy of each blocked dipeptide was included, corresponding to a concentration of 0.012 M for each salt-bridging partner, with a distance of 10 Å between the amino acids. All force field parameters of the side-chain analogues were based on those of the complete amino acids. For the chloride ions, parameters derived specifically for the water model were used when available; otherwise, parameters derived for a similar water model were used.[53–55] Nonbonded parameters of the side-chain analogues, along with those used to model chloride ions and blocked amino acid dipeptides are provided in the associated publication.[56]

To alleviate any unfavorable interactions, each model was subjected to energy minimization followed by a two-stage equilibration with harmonic position restraints on all heavy atoms of the side-chain analogues (force constant of 10 kcal mol^{-1} $Å^{-2}$) using the Desmond 3.0.1.0 software package.[57] In the first stage, the energy-minimized system was equilibrated for 20 ps at

constant temperature (25 °C) using a weak Langevin thermostat[58] (frictional constant of 1 ps^{-1}). During the second stage, the system was equilibrated for 1 ns at constant temperature (25 °C) and pressure (1 atm) using the Martyna-Tobias-Klein thermostat and barostat[59] (coupling time constants of 1.0 ps and 2.0 ps, respectively). To enable a 2-fs time step, bonds to hydrogen were constrained to their equilibrium values using the M-SHAKE algorithm.[60] A short-range nonbonded cutoff of 10.0 Å was used, and long-range electrostatics were calculated using the particle mesh Ewald (PME) method.[61] The frame from the second half of the NPT equilibration with volume closest to the average was used to start the production simulation.

2.2.2 Simulation details

To obtain extensive sampling of salt bridge association (and dissociation) events, 1-μs MD simulations were performed for each side-chain analogue system; 10-μs simulations were performed for the blocked arginine and aspartate dipeptide systems. All simulations were carried out in the NVT ensemble using a 64-node Anton special-purpose supercomputer, which is able to run MD simulations roughly two orders of magnitude faster than conventional hardware[25] (altogether, the simulations required a total of 40 machine-days). The temperature was maintained at 25 °C using the Nosé-Hoover thermostat[62] with a weak coupling constant of 0.5 ps. Van der Waals and short-range electrostatic interactions were truncated at 10.0 Å; long-range electrostatic interactions were calculated using the Gaussian split Ewald method.[63] To enable a 2.5-fs time step, bonds to hydrogen were constrained to their equilibrium lengths using the M-SHAKE algorithm.[60] Conformations were saved every picosecond for analysis.

2.2.3 Calculation of equilibrium association constants

Equilibrium association constants (K_A) were calculated from the populations of the bound and unbound states of the oppositely charged side-chain analogues. For example, the K_A for association between guanidinium and acetate was calculated using the following:

$$K_A = \frac{[bound\ state]}{[unbound\ state][unbound\ guanidinium]} = \left(\frac{P_{bound}}{P_{unbound\ acetate} P_{unbound\ guanidinium}}\right)\left(\frac{1}{C_0}\right) \qquad (2.1)$$

where P_{bound} is the population of the bound state; $P_{unbound\ guanidinium}$ and $P_{unbound\ acetate}$ are the populations of unbound guanidinium and acetate, respectively; and C_0 is the reference concentration of guanidinium (i.e., 0.9 M). In addition to species in which a single acetate molecule is bound to a single cation molecule, forming a 1:1 complex (e.g., the guanidinium/acetate complex), species in which acetate is bound to two cation molecules, forming a 1:2 complex (e.g., the diguanidinium/acetate complex) were observed. K_A values for the latter are included in Table 2.3; the results discussed below focus on formation of the major complex, which is the 1:1 complex. Standard errors in the K_A values were calculated using a block averaging method.[64]

For each side-chain analogue system, the unbound and bound states were defined using the potential of mean force (PMF) as a function of the minimum distance between the nitrogen and oxygen atoms of the positively and negatively charged analogues, respectively (minimum N-O distance; see Figure 2.1). In particular, the point of inflection between the bound state free energy minimum (~2.5-3 Å) and the desolvation barrier (~3-3.5 Å) was used as the bound state cutoff, while 4.5 Å was used as the unbound state cutoff. If the minimum N-O distance between an analogue pair dropped below the bound state cutoff they were classified as bound until they crossed the unbound state cutoff, and vice versa. For simulations of the blocked arginine and aspartate

dipeptides, the same definitions of the unbound and bound states were used as for the guanidinium/acetate system.

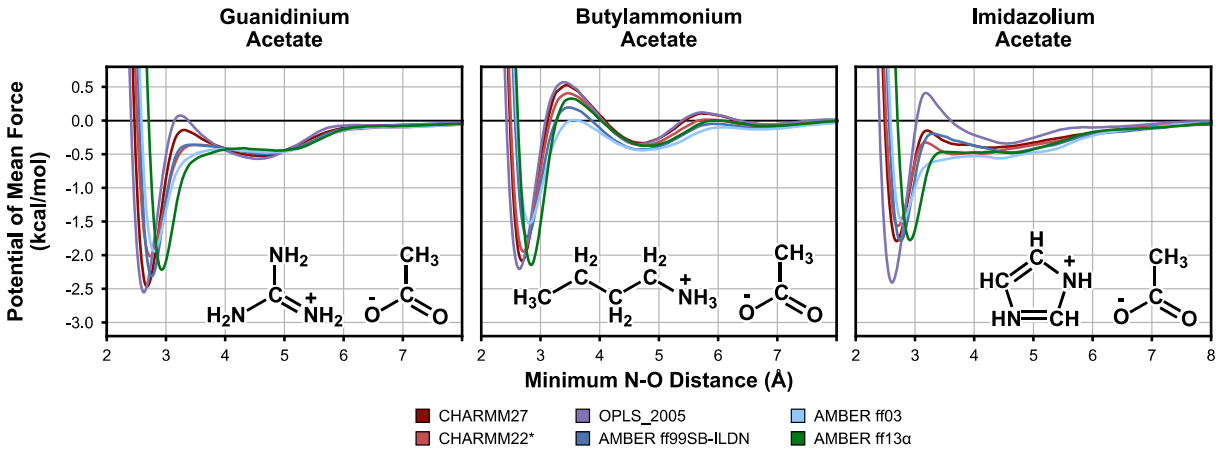

Figure 2.1. Potentials of mean force (PMF) between three different pairs of oppositely charged side-chain analogues using six biomolecular force fields with the TIP4P-Ew explicit water model.

2.2.4 Calculation of the solvent dielectric constant

The dielectric constant of water in each simulation, ε_{water}, was calculated using the following equation:

$$\varepsilon = 1 + \frac{\langle M_{water}^2 \rangle - \langle M_{water} \rangle^2}{3\varepsilon_0 V_{water} k_B T} \tag{2.2}$$

where M_{water} is the net dipole moment of water, V_{water} is the volume occupied by water, T is the temperature of the system, k_B is the Boltzmann constant, and ε_0 is the permittivity of free space. The net dipole moment of water was calculated using the following:

$$M_{water} = \sum_{i=1}^{N} q_i r_i \tag{2.3}$$

where q is the atomic charge of each water site and r is its position vector.[65,66] The dielectric constant of water, rather than that of the complete system, was used since it is impossible to calculate the contributions of molecules with a net charge to the system dipole moment from simulations with periodic boundary conditions.[66] The appropriate volume was thus the volume of the water molecules present in the system. For each water model used, a pure water system of the same total volume as that of the amino acid analogue systems was equilibrated using the same protocol and the molecular volume of water calculated. For each analogue system, the number of water molecules present was multiplied by the molecular volume to calculate the approximate volume of water present in the system. Standard errors in the ε_{water} values were calculated using a block averaging method.[64]

2.3 RESULTS AND DISCUSSION

We compared the following six current biomolecular force fields in terms of their ability to model salt bridge interactions: AMBER ff99SB-ILDN,[28] AMBER ff03,[31] AMBER ff13α,[33] CHARMM27,[67] CHARMM22*,[32] and OPLS_2005.[68] In particular, we simulated association (and dissociation) of salt bridges between the following three pairs of oppositely charged amino acids: Arg/Asp, Lys/Asp, and His(+)/Asp. We focused primarily on simulating side-chain analogues (i.e., guanidinium, butylammonium, and imidazolium cations for arginine, lysine, and histidine, respectively, and acetate anion for aspartate) since these analogues are the minimal systems for studying the formation of salt bridges. In addition, equilibrium association constant (K_A) values for such systems have been experimentally measured,[51,69] providing an excellent opportunity to validate the simulations. While blocked amino acid dipeptides (i.e., acetyl—amino acid—N-

methylamide) might be regarded as being more representative of the protein environment, no experimental K_A values for the association of oppositely charged amino acid dipeptides are available. Nonetheless, we evaluated the force fields in simulating such systems, focusing on just one of the three salt bridges, Arg/Asp. Finally, in addition to the above simulations, in which each biomolecular force field was paired with the TIP4P-Ew explicit water model, which reproduces the liquid properties of water at the temperatures and pressures relevant to biology,[27] we also evaluated the influence of the water model on the strength of the salt bridges by testing a selection of force field/water model combinations for all three pairs of side-chain analogues. For each force field a selection of water models drawn from TIP3P,[70] mTIP3P,[71] TIP4P,[70] TIP4P/2005,[72] and SPC/E[73] were tested, including the water model with which each force field was originally derived.

2.3.1 Association constants of side-chain analogues

To validate our simulations of association between oppositely charged side-chain analogues, we computed K_A values and compared these to those measured by experiments. Experimental K_A values have been measured for guanidinium/acetate and butylammonium/acetate association by monitoring changes in the pK_a of acetate in the presence or absence of either the guanidinium or butylammonium cation.[51] Our microsecond-long simulations yielded thousands of independent binding events, permitting the extraction of extremely precise K_A values, with the mean lifetimes of the bound state ranging from ~10-300 ps and the mean lifetimes of the unbound state ranging from ~20-120 ps (Table 2.3).

In general, the K_A values computed from our side-chain analogue simulations are overestimated in comparison to experimentally measured values, with the AMBER ff03 force field overestimating the strengths of the salt bridges to the least extent and the OPLS_2005 force field

to the greatest extent, when using the same water model (Table 2.1). The computed K_A values vary considerably among the force fields. For example, when using the TIP4P-Ew water model, the computed K_A values for the three types of salt bridges vary by as much as ~4-fold, ~3-fold, and ~4-fold for the associations of guanidinium, butylammonium, and imidazolium with acetate, respectively, which amounts to ~1.4-fold, ~1.8-fold, and ~1.9-fold differences in the probabilities of binding (P_{bound}) (see Figure 2.2). We note that our definition of the bound state is very conservative and that the use of less conservative definitions (e.g., use of the desolvation barrier as a cutoff) yields even stronger association constants, without affecting our overall conclusions.

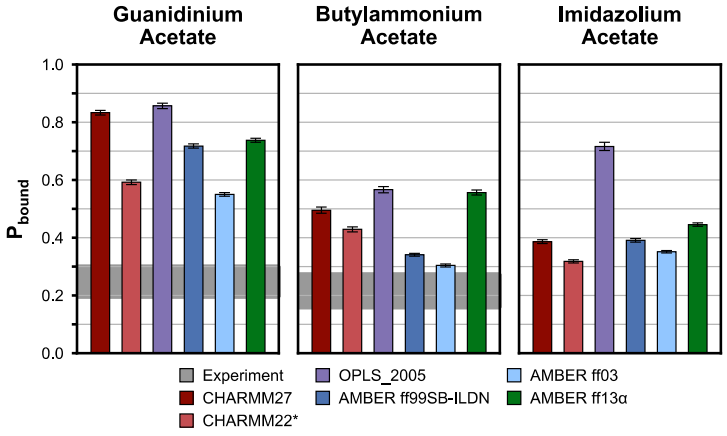

Figure 2.2. Probabilities of binding (P_{bound}) between acetate and one or more molecules of three cationic side-chain analogues using six biomolecular force fields with the TIP4P-Ew explicit water model. The P_{bound} values that correspond to the experimentally determined K_A values of guanidinium acetate and butylammonium acetate are depicted as horizontal gray bars;[51] no experimentally measured K_A is available for the imidazolium acetate system. Error bars represent 95% confidence intervals calculated using a block averaging method.[64]

Table 2.1. Association constants (K_A) and probabilities of binding (P_{bound}) for three different pairs of oppositely charged side-chain analogues using six biomolecular force fields and six explicit water models. Results are from 1-μs simulations and standard errors were calculated using a block averaging method.[64]

Force Field	Water Model	Guanidinium/Acetate		Butylammonium/Acetate		Imidazolium/Acetate	
		P_{bound}[a]	K_A (M^{-1})	P_{bound}[a]	K_A (M^{-1})	P_{bound}[a]	K_A (M^{-1})
AMBER ff99SB-ILDN	TIP4P-Ew	0.57	2.23 ± 0.03	0.32	0.53 ± 0.01	0.36	0.65 ± 0.01
AMBER ff03	TIP4P-Ew	0.45	1.12 ± 0.01	0.28	0.45 ± 0.00	0.32	0.54 ± 0.00
AMBER ff13ALPHA	TIP4P-Ew	0.54	2.28 ± 0.03	0.48	1.20 ± 0.01	0.40	0.79 ± 0.01
CHARMM27	TIP4P-Ew	0.61	4.06 ± 0.10	0.45	0.98 ± 0.01	0.35	0.63 ± 0.01
CHARMM22*	TIP4P-Ew	0.48	1.31 ± 0.02	0.39	0.75 ± 0.01	0.29	0.47 ± 0.00
OPLS-AA	TIP4P-Ew	0.63	5.03 ± 0.25	0.53	1.43 ± 0.03	0.57	2.25 ± 0.06
OPLS_2005	TIP4P-Ew	0.63	4.92 ± 0.17	0.50	1.27 ± 0.02	0.55	1.96 ± 0.04
AMBER ff99SB-ILDN	TIP3P	0.57	4.52 ± 0.09	0.39	0.78 ± 0.01	0.45	1.04 ± 0.01
AMBER ff03	TIP3P	0.49	1.53 ± 0.01	0.33	0.58 ± 0.00	0.36	0.65 ± 0.00
CHARMM27	TIP3P	0.53	9.03 ± 0.34	0.52	1.65 ± 0.03	0.44	1.00 ± 0.01
CHARMM22*	TIP3P	0.52	1.88 ± 0.02	0.44	1.03 ± 0.01	0.34	0.59 ± 0.00
OPLS-AA	TIP3P	0.56	13.57 ± 0.73	0.55	2.42 ± 0.05	0.55	3.85 ± 0.12
OPLS_2005	TIP3P	0.59	11.65 ± 0.59	0.54	2.22 ± 0.04	0.55	3.37 ± 0.10
CHARMM27	mTIP3P	0.58	6.34 ± 0.20	0.47	1.20 ± 0.02	0.39	0.77 ± 0.01
CHARMM22*	mTIP3P	0.49	1.41 ± 0.01	0.50	0.76 ± 0.01	0.29	0.48 ± 0.00
CHARMM22*	mTIP3P	0.49	1.38 ± 0.01				
OPLS_2005	TIP4P	0.60	8.18 ± 0.30			0.57	3.07 ± 0.07
AMBER ff03	TIP4P/2005	0.42	0.94 ± 0.01			0.29	0.46 ± 0.00
AMBER ff99SB-ILDN	SPC/E	0.60	2.98 ± 0.05			0.42	0.84 ± 0.01
CHARMM27	SPC/E	0.64	5.62 ± 0.17	0.49	1.23 ± 0.02	0.40	0.78 ± 0.01
CHARMM22*	SPC/E	0.51	1.54 ± 0.02			0.31	0.52 ± 0.00
OPLS_2005	SPC/E	0.66	5.26 ± 0.19			0.58	2.13 ± 0.04
Experiment[51,69,][b]		~0.25 ± 0.03	~0.37 ± 0.05	~0.22 ± 0.03	~0.31 ± 0.05		

[a] Standard errors of P_{bound} were uniformly ≤0.01.

[b] Experimental K_A values of guanidinium and butylammonium acetate permit only a qualitative estimate of the associated error. Taking two experimentally measured K_A values of guanidinium acetate using different protocols into account,[51,69] we estimate an error of ±0.05, although the true uncertainty is not known. Using this estimate, we have back-calculated the range of simulated P_{bound} values that would be expected in our simulation, based on the experimental K_A.

One potential factor that could influence the degree of salt bridge formation in our simulations is the choice of force field parameters for the chloride ions. To verify that these parameters are not the cause for overestimating salt bridge strength, we carried out simulations of a single guanidinium/acetate pair (corresponding to concentrations of 0.1 M) with no chloride ions present. The resulting K_A values are even higher than those measured in the presence of chloride ions, indicating that the chloride ion parameters do not cause disproportionate salt bridge stability (Table 2.4).

Two obvious features of a force field that influence the strength of the salt bridges are the atomic charges and radii. As expected, the CHARMM22* force field yields K_A values that are closer to experiment than those from the parent CHARMM27 force field since the atomic charges for the arginine, aspartate, and glutamate residues were parametrized specifically to reproduce the experimental association of guanidinium acetate.[32] However, the CHARMM22* force field does not produce as close agreement with experiment as the AMBER ff03 force field, which shows good agreement for butylammonium acetate. Given these results, it appears that the general strategy used to derive atomic charges for the AMBER ff03 force field is reasonably effective for modeling electrostatic interactions. This strategy involved the derivation of atomic charges in the presence of a continuum solvent model with a dielectric constant of 4 to mimic an organic solvent (protein-like) environment. The resulting atomic charges in the AMBER ff03 force field are notably less polarized than those in the AMBER ff9X family (including the AMBER ff99SB-ILDN force field tested here), which were derived in vacuum and share the same set of atomic radii. The AMBER ff03 atomic charges are also less polarized than those in the AMBER ff13α force field, with atomic charges possessing increased polarity relative to previous AMBER charge models.[33] In the AMBER ff13α charge model, nonpolarizable point charges have been fit to

implicitly account for solvent polarization, using iterative cycles of classical MD simulations with explicit water (i.e., TIP4P-Ew) to estimate the water charge density around the solute, followed by quantum mechanical calculations to determine updated solute charges.

Interestingly, although certain critical atomic radii (e.g., the nitrogen in butylammonium and the oxygen in acetate) in the AMBER ff13α force field were adjusted from their original values in the AMBER ff99 force field to reproduce experimental hydration free energies of the relevant amino acid analogues,[33] the resulting strengths of the salt bridges are more overestimated, relative to the other tested AMBER force fields. Notably, the AMBER ff13α force field results in a free energy landscape for salt bridge formation that is significantly different from those of the other force fields. In particular, as shown by the PMFs as a function of the minimum N-O distance between the oppositely charged analogues for the three types of salt bridges (Figure 2.1), the free energy minima for the bound states are consistently shifted to the right in the AMBER ff13α force field relative to the other force fields. When we substituted the atomic radii in the AMBER ff13α force field with the original radii from the AMBER ff99 force field, the free energy minima for the bound states shifted back towards those of the other force fields and yielded significantly deeper minima as well as more pronounced desolvation barriers, particularly for the guanidinium/acetate and imidazolium/acetate systems (Figure 2.7).

Since the atomic charges of the OPLS_2005 force field are not significantly different from those of the other force fields, the most likely reason for the fact that this force field overestimates the K_A values to the greatest extent is that the atomic radii of the nitrogen-attached hydrogen atoms are smaller than those used by the other force fields, potentially allowing the pairs to associate more closely and increasing their electrostatic attraction. Consistent with this notion, simulations using the OPLS-AA force field, which differs from the OPLS_2005 force field only in that it omits

atomic radii for these hydrogen atoms, resulted in slightly more strongly associated salt bridges (Table 2.3). We note that the ranking of the strengths of the three types of side-chain salt bridges in our study by the AMBER ff99SB-ILDN, CHARMM27, and OPLS-AA force fields is consistent with that observed for their oppositely charged termini in a recent study by others.[50]

2.3.2 Association constants of amino acid dipeptides

As mentioned above, we additionally tested salt bridge formation of the Arg/Asp pair by simulating association/dissociation of blocked amino acid dipeptides, testing six different force fields in conjunction with the TIP4P-Ew explicit water model. As shown in Table 2.2, the relative ranking of the force fields in terms of the K_A is generally consistent with our results from the corresponding side-chain analogue system (guanidinium/acetate). The only exception is the AMBER ff13α force field, which yields the weakest K_A for the association of the amino acid dipeptides, as opposed to an intermediate K_A value for the association of guanidinium/acetate. As indicated by the PMF between the arginine and aspartate dipeptides (Figure 2.3), the bound state free energy minimum of the AMBER ff13α force field is the most shallow among the tested force fields, corresponding to the lowest frequency of salt bridge formation. The inclusion of the backbone groups, therefore, appears to alter its propensity for salt bridge formation, likely through the competition of side-chain/backbone interactions with the side-chain/side-chain interactions between the two amino acids. This result emphasizes the benefit of using unbiased simulations; had the relative orientations of the amino acids been fixed as in previous studies,[14] any effects of significant side-chain/backbone interactions on the frequency of salt bridge formation would not have been apparent.

Table 2.2. Association constants (K_A) and probabilities of side-chain/side-chain ($P_{SC/SC\ bound}$) and side-chain/backbone association ($P_{SC/BB\ bound}$), respectively, for blocked arginine and aspartate dipeptides using six biomolecular force fields with the TIP4P-Ew water model. Results are from 10-μs simulations and standard errors were calculated using a block averaging method.[64]

Force Field	Water Model	Arginine/Aspartate			
		$P_{SC/SC\ bound}$	K_A (M^{-1})	$P_{SC/BB\ bound}$	$P_{SC/SC\ bound}$ / $P_{SC/BB\ bound}$
AMBER ff99SB-ILDN	TIP4P-Ew	0.041 ± 0.003	3.51 ± 0.94	0.010 ± 0.001	4.1 ± 0.4
AMBER ff03	TIP4P-Ew	0.033 ± 0.002	3.24 ± 0.67	0.017 ± 0.001	1.9 ± 0.2
AMBER ff13α	TIP4P-Ew	0.023 ± 0.002	1.83 ± 0.43	0.018 ± 0.001	1.3 ± 0.1
CHARMM27	TIP4P-Ew	0.113 ± 0.012	7.53 ± 3.60	0.007 ± 0.001	16.4 ± 2.0
CHARMM22*	TIP4P-Ew	0.038 ± 0.003	3.11 ± 0.65	0.007 ± 0.001	5.5 ± 0.5
OPLS_2005	TIP4P-Ew	0.153 ± 0.016	19.07 ± 8.58	0.013 ± 0.001	11.8 ± 1.6

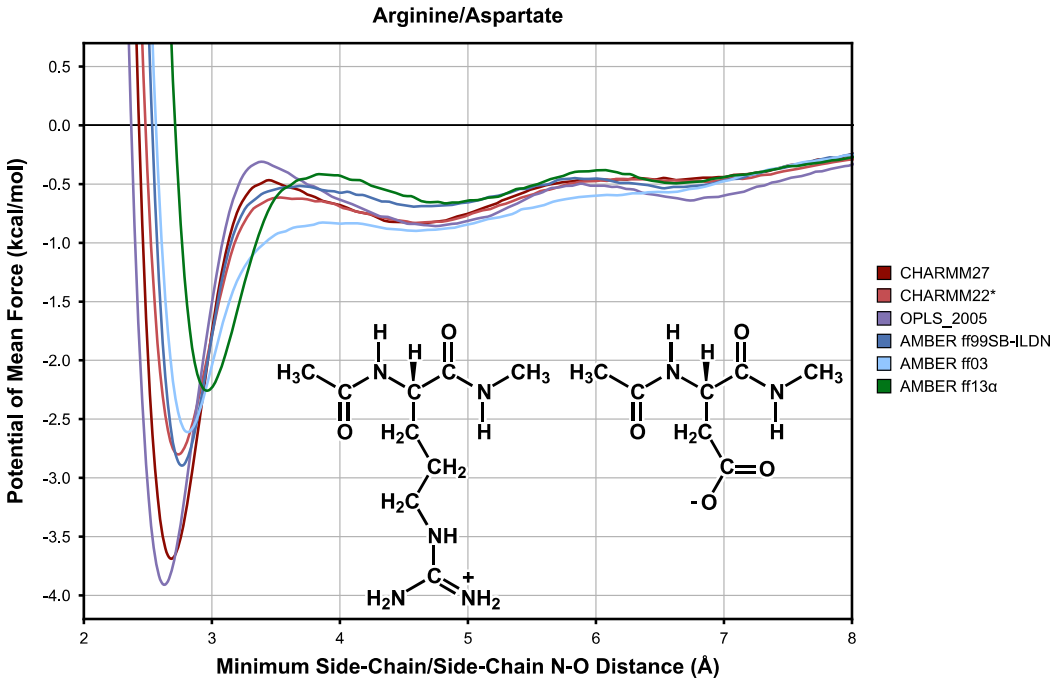

Figure 2.3. Potentials of mean force (PMF) between blocked arginine and aspartate dipeptides using six biomolecular force fields with the TIP4P-Ew water model. The larger noise level compared to the data presented in **Figure 2.1** is caused by simulating a single pair of binding partners, rather than a concentrated solution.

To monitor side-chain/backbone association, we used the same minimum N-O distance coordinate and bound and unbound state definitions as used for the side-chain/side-chain interactions. A comparison of the relative probabilities of side-chain/side-chain versus side-

chain/backbone association (Figure 2.4) reveals that the force fields generally prefer side-chain/side-chain association by a factor of ~2x or more over side-chain/backbone association. The exception is AMBER ff13α, which shows a lower preference for side-chain/side-chain association of ~1.3x. This slight preference over side-chain/backbone association is likely due to the substantially more polarized backbone amide and carbonyl groups of the AMBER ff13α force field relative to previous AMBER force fields (including the AMBER ff99SB-ILDN and AMBER ff03 force fields).[33] Thus, as a result of the delicate balance of side-chain/side-chain and side-chain/backbone interactions, the strength of the Arg/Asp salt bridge appears to be most accurately modeled (least overstabilized) by the AMBER ff13α force field in a model system that is representative of a protein environment.

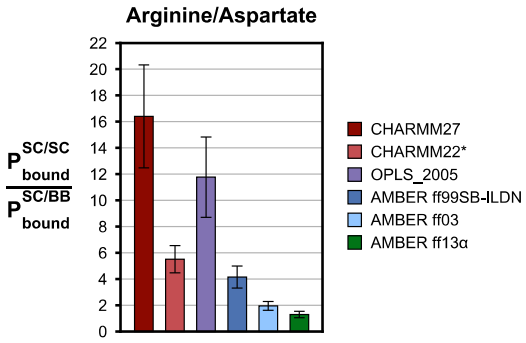

Figure 2.4. Probability of side-chain/side-chain association $\left(\mathbf{P}^{\mathbf{SC/SC}}_{\mathbf{bound}}\right)$ over the probability of side-chain/backbone association $\left(\mathbf{P}^{\mathbf{SC/BB}}_{\mathbf{bound}}\right)$ between blocked arginine and aspartate dipeptides using six biomolecular force fields with the TIP4P-Ew water model. Error bars represent 95% confidence intervals calculated using a block averaging method.[64]

2.3.3 Influence of the water model

In addition to the force field, the choice of water model can affect the strength of salt bridges. To evaluate the influence of the water model, we tested the above three side-chain analogue systems with a selection of force field/water model combinations in addition to the force field/TIP4P-Ew combinations. Regardless of the water

model, the relative ranking of the force fields is unchanged in terms of the K_A values, with P_{bound} varying by 5-10% between the water models (Table 2.1). We also evaluated the dependence of salt bridge interactions on the dielectric constant of the employed water model (ε_{water}). Interestingly, despite the fact that the SPC/E water model yields a computed dielectric constant (ε_{water} = 70;

Table 2.5) that is closest to the experimental value (ε_{water} = 78.4)[74] among all of the tested water models, the use of the SPC/E water model results in stronger salt bridge interactions than seen with the TIP4P-Ew water model, in which the ε_{water} value is underestimated (ε_{water}= 56;

Table 2.5). In fact, as shown in Figure 2.5, there appears to be no clear correlation between ε_{water} of the water model and the strength of the salt bridges. As an aside, the CHARMM27 and CHARMM22* force fields were tested with both the standard TIP3P and the CHARMM-modified TIP3P (mTIP3P) water model with which they were developed. The mTIP3P water model includes atomic radii on hydrogen as well as oxygen atoms,[71] whereas standard TIP3P includes only the oxygen atom. Using the mTIP3P water model consistently results in lower K_A values, in better agreement with experiment. This suggests that it may be advisable to use of the CHARMM-modified TIP3P, rather than the standard TIP3P water model with any CHARMM force field.

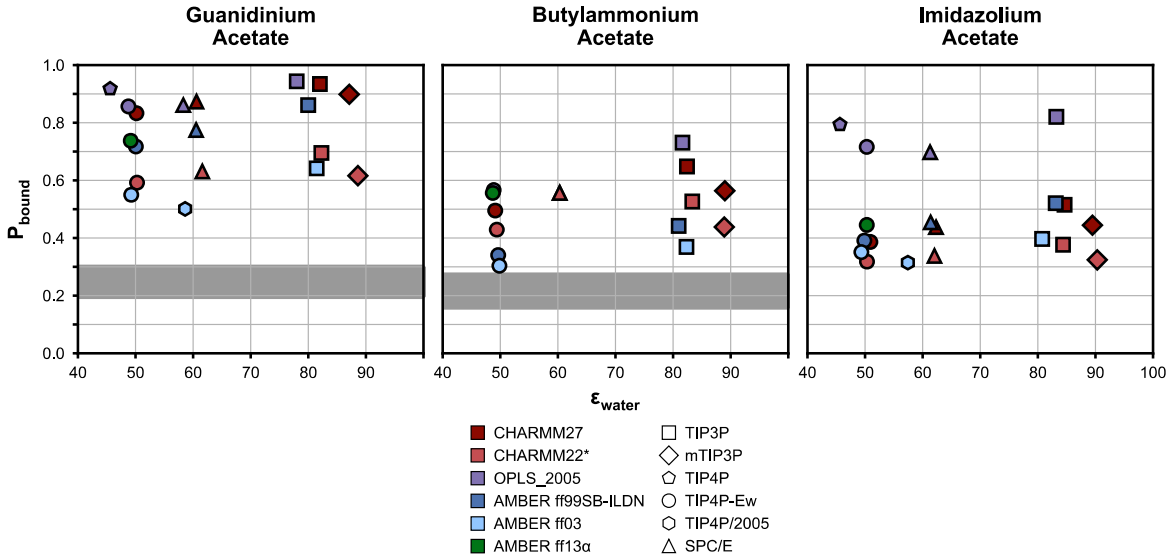

Figure 2.5. Relationship between the probability of binding (P_{bound}) and the dielectric constant of the water model (ε_{water}) for three different pairs of oppositely charged side-chain analogues simulated using six biomolecular force

fields and six explicit water models. The P_{bound} values that correspond to the experimental association constants (K_A) for the guanidinium acetate and butylammonium acetate systems are depicted as horizontal gray bars;[51,69] no experimentally measured K_A is available for the imidazolium acetate system. The ε_{water} values were calculated from the first 100 ns of each simulation. For each model, the presence of the solutes lowers the ε_{water} for each system by 10−15 relative to that of pure water (

Table 2.5). Note that the error bars are not visible, because 95% confidence intervals for both P_{bound} and ε_{water} lie within the symbols' area in the graph.

2.4 CONCLUSIONS

We compared the modeling of salt bridge interactions using six current biomolecular force fields. Three different salt bridges (Arg/Asp, Lys/Asp, and His(+)/Asp) were simulated and considerable differences in their strengths were noted, both between the force fields and with experiment. Given the availability of experimentally measured K_A values for the association of oppositely charged side-chain analogues, we have focused primarily on modeling salt bridge formation using these systems. We also tested the applicability of our results to amino acids by simulating blocked amino acid dipeptides for one of the salt bridges, Arg/Asp.

Our side-chain analogue simulations reveal that the computed K_A values are generally overestimated, relative to experimental values, with the AMBER ff03 force field overestimating the strengths of the salt bridges to the least extent and the OPLS_2005 force field to the greatest extent when using the same water model (TIP4P-Ew). For the blocked arginine and aspartate dipeptides, we observed general agreement in the relative ranking of the force fields with that obtained from simulations with the corresponding side-chain analogues. The only exception is the AMBER ff13α force field, which resulted in the lowest probability of salt bridge formation, likely

due to the presence of competing side-chain/backbone interactions. Thus, while the AMBER ff03 force field overestimates the strength of salt bridges to the least extent for the side-chain analogue systems, the AMBER ff13α force field results in an even lower frequency of salt bridge formation than the AMBER ff03 force field for the complete amino acids. Finally, we examined the influence of the water model on the strengths of the salt bridges. Irrespective of the water model, the relative ranking of the force fields remained unchanged, with no clear correlation between the probability of binding (salt bridge formation) and the dielectric constant of the solvent (ε_{water}).

In conclusion, when running MD simulations in which salt bridge formation may be of interest, careful attention should be paid to the specific force field and water model in MD simulations of protein systems. Several current force fields yield considerably higher K_A values than those experimentally determined, a discrepancy that may lead to erroneous conclusions. Our encouraging results with the AMBER ff13α force field suggest that charge derivation strategies which implicitly incorporate solvent polarization from explicit water may significantly extend the lifetime of fixed-charge force fields, which include all of the force fields tested in this study. Nonetheless, departures from these simple point charge models may also be necessary. For example, using polarizable force fields that permit varying the charge distribution within a molecule based on both its conformation and environment may alleviate such shortcomings. In the past, the solvation of ions[75,76] and charged small molecules,[77] have been modeled using polarizable force fields, resulting in improved agreement with experiment, compared to CHARMM27 and AMBER ff99 force fields (equivalent to the AMBER ff99SB-ILDN force field used here). Future work will determine whether or not this also holds for protein salt bridges

2.5 ACKNOWLEDGEMENTS

This work was supported by start-up funds from the University of Pittsburgh School of Arts & Sciences to LTC and start-up funds from the University of Pittsburgh School of Medicine to AMG. KTD was supported by National Institutes of Health training grant GM088119. Anton computer time was provided by the National Resource for Biomedical Supercomputing (NRBSC), the Pittsburgh Supercomputing Center (PSC), and the BTRC for Multiscale Modeling of Biological Systems (MMBioS) through Grant P41GM103712-S1 from the National Institutes of Health. The Anton computer at NRBSC/PSC was generously made available by D. E. Shaw Research.

2.6 SUPPORTING INFORMATION

2.6.1 Supporting figures

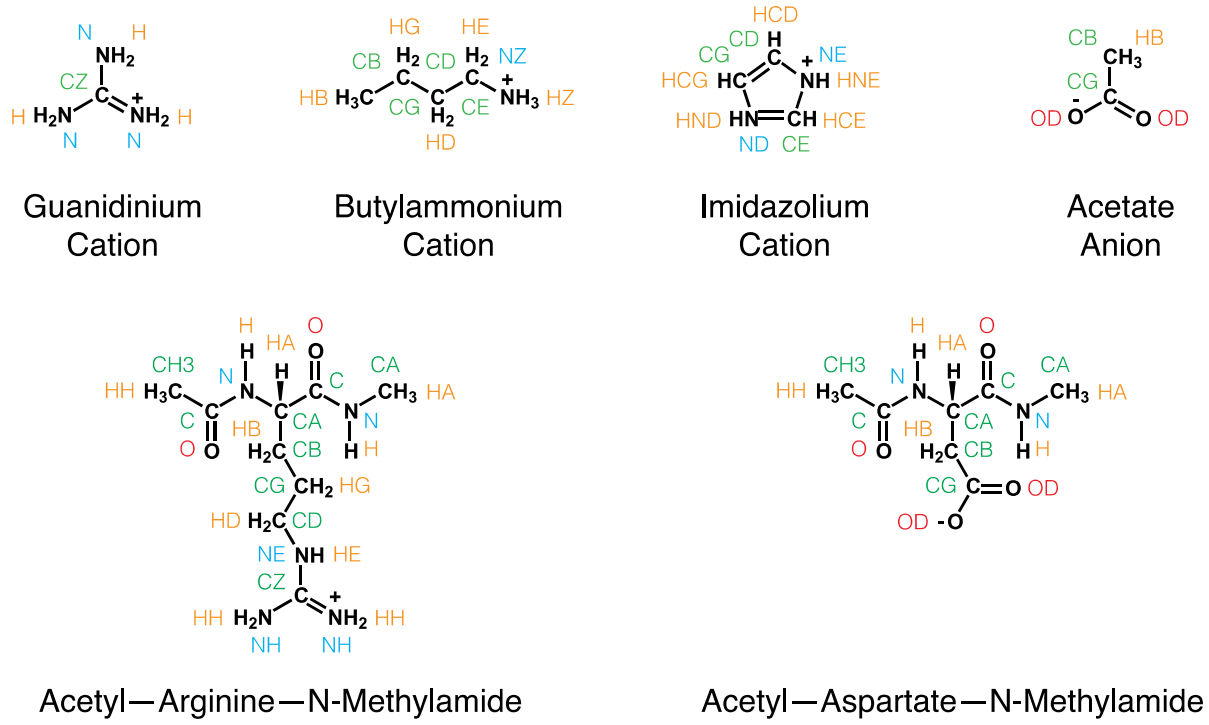

Figure 2.6. Structures and atom names for side-chain analogues and blocked amino acid dipeptides.

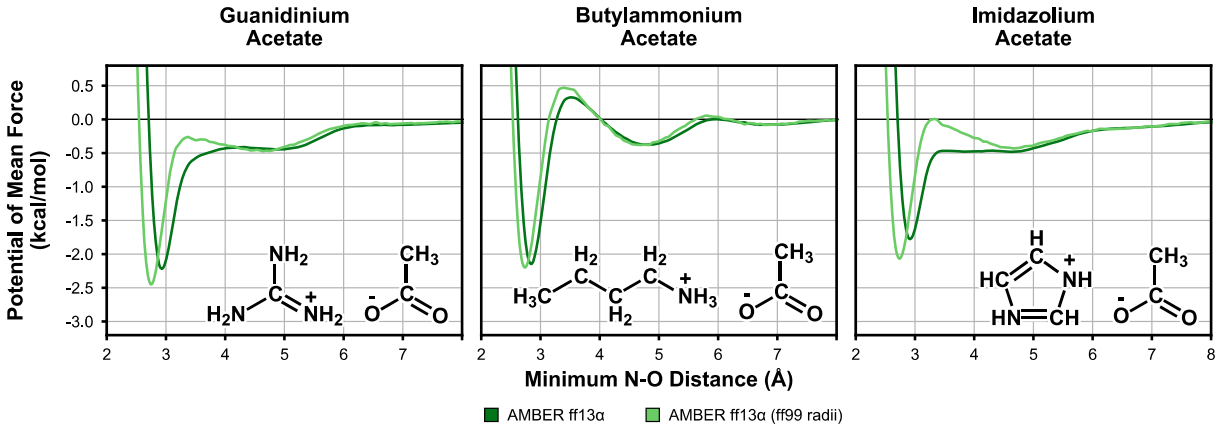

Figure 2.7. Potentials of mean force (PMF) between three different pairs of oppositely charged amino acid analogues using the AMBER ff13α force field with its modified atomic radii or with those of the AMBER ff99 force field. The former simulation was 1 μs in duration while the latter was 100 ns in duration.

2.6.2 Supporting tables

Table 2.3. Association constants (K_A), probabilities of binding (P_{bound}), and mean lifetimes of the bound and unbound states for three different pairs of oppositely charged side-chain analogues using six biomolecular force fields and six explicit water models. In addition, the OPLS-AA force field, which differs from the OPLS_2005 force field only in the omission of atomic radii on nitrogen-attached hydrogens, was tested. Results are from 1-μs simulations (with the exception of those run with the OPLS-AA force field and TIP4P-Ew water model, which were 500-ns in duration) and standard errors were calculated using a block averaging method.[64]

Guanidinium (G) / Acetate (A)

Force Field	Water Model	G + A ⇌ GA		G + GA ⇌ G_2A		Mean Lifetime of Bound State (ps)[c]	Mean Lifetime of Unbound State (ps)[c]
		P_{bound}[c]	K_A (M^{-1})	P_{bound}[c]	K_A (M^{-1})[c]		
AMBER ff99SB-ILDN	TIP4P-Ew	0.57	2.23 ± 0.03	0.14	0.28	46	34
AMBER ff03	TIP4P-Ew	0.45	1.12 ± 0.01	0.09	0.22	22	25
AMBER ff13α	TIP4P-Ew	0.54	2.28 ± 0.03	0.18	0.37	30	22
CHARMM27	TIP4P-Ew	0.61	4.06 ± 0.10	0.21	0.39	82	45
CHARMM22*	TIP4P-Ew	0.48	1.31 ± 0.02	0.10	0.24	31	33
OPLS-AA	TIP4P-Ew[d]	0.63	5.03 ± 0.25	0.22	0.40	169	75
OPLS_2005	TIP4P-Ew	0.63	4.92 ± 0.17	0.21	0.38	161	76
AMBER ff99SB-ILDN	TIP3P	0.57	4.52 ± 0.09	0.26	0.52	53	28
AMBER ff03	TIP3P	0.49	1.53 ± 0.01	0.14	0.31	24	22
CHARMM27	TIP3P	0.53	9.03 ± 0.34	0.36	0.75	98	43
CHARMM22*	TIP3P	0.52	1.88 ± 0.02	0.16	0.34	35	29
OPLS-AA	TIP3P[d]	0.56	13.57 ± 0.73	0.36	0.71	232	75
OPLS_2005	TIP3P	0.59	11.65 ± 0.59	0.32	0.61	214	75
CHARMM27	mTIP3P	0.58	6.34 ± 0.20	0.29	0.56	95	48
CHARMM22*	mTIP3P	0.49	1.41 ± 0.01	0.12	0.27	32	33
CHARMM22*	mTIP3P	0.49	1.38 ± 0.01	0.11	0.25	32	33
OPLS_2005	TIP4P	0.60	8.18 ± 0.30	0.29	0.54	144	49
AMBER ff03	TIP4P/2005	0.42	0.94 ± 0.01	0.07	0.19	21	27
AMBER ff99SB-ILDN	SPC/E	0.60	2.98 ± 0.05	0.16	0.30	54	35
CHARMM27	SPC/E	0.64	5.62 ± 0.17	0.22	0.38	96	46
CHARMM22*	SPC/E	0.51	1.54 ± 0.02	0.11	0.24	36	34
OPLS_2005	SPC/E	0.66	5.26 ± 0.19	0.19	0.33	192	89
Experiment[51,69,e]		~0.25 ± 0.03	~0.37 ± 0.05				

Butylammonium (B) / Acetate (A)

Force Field	Water Model	B + A ⇌ BA		B + BA ⇌ B_2A		Mean Lifetime of Bound State (ps)[c]	Mean Lifetime of Unbound State (ps)[c]
		P_{bound}[c]	K_A (M^{-1})	P_{bound}[c]	K_A (M^{-1})[c]		
AMBER ff99SB-ILDN	TIP4P-Ew	0.32	0.53 ± 0.01	0.02	0.07	27	58
AMBER ff03	TIP4P-Ew	0.28	0.45 ± 0.00	0.02	0.07	16	30
AMBER ff13α	TIP4P-Ew	0.48	1.20 ± 0.01	0.07	0.16	53	53
CHARMM27	TIP4P-Ew	0.45	0.98 ± 0.01	0.04	0.10	78	91
CHARMM22*	TIP4P-Ew	0.39	0.75 ± 0.01	0.04	0.11	48	74
OPLS-AA	TIP4P-Ew[d]	0.53	1.43 ± 0.03	0.07	0.15	103	86
OPLS_2005	TIP4P-Ew	0.50	1.27 ± 0.02	0.06	0.14	99	91
AMBER ff99SB-ILDN	TIP3P	0.39	0.78 ± 0.01	0.05	0.13	34	50
AMBER ff03	TIP3P	0.33	0.58 ± 0.00	0.03	0.12	17	33
CHARMM27	TIP3P	0.52	1.65 ± 0.03	0.12	0.26	127	89
CHARMM22*	TIP3P	0.44	1.03 ± 0.01	0.08	0.20	68	74
OPLS-AA	TIP3P[d]	0.55	2.42 ± 0.05	0.18	0.37	172	85
OPLS_2005	TIP3P	0.54	2.22 ± 0.04	0.18	0.37	164	86
CHARMM27	mTIP3P	0.47	1.20 ± 0.02	0.09	0.20	112	105
CHARMM22*	mTIP3P	0.50	0.76 ± 0.01	0.06	0.14	57	87
CHARMM27	SPC/E	0.49	1.23 ± 0.02	0.06	0.14	107	102
Experiment[51,69,e]		~0.22 ± 0.03	~0.31 ± 0.05				

Imidazolium (I) / Acetate (A)

Force Field	Water Model	I + A ⇌ IA		I + IA ⇌ I_2A		Mean Lifetime of Bound State (ps)[c]	Mean Lifetime of Unbound State (ps)[c]
		P_{bound}[c]	K_A (M^{-1})	P_{bound}[c]	K_A (M^{-1})[c]		
AMBER ff99SB-ILDN	TIP4P-Ew	0.36	0.65 ± 0.01	0.03	0.10	23	43
AMBER ff03	TIP4P-Ew	0.32	0.54 ± 0.00	0.03	0.10	13	27
AMBER ff13α	TIP4P-Ew	0.40	0.79 ± 0.01	0.04	0.12	18	28
CHARMM27	TIP4P-Ew	0.35	0.63 ± 0.01	0.03	0.10	37	67
CHARMM22*	TIP4P-Ew	0.29	0.47 ± 0.00	0.02	0.09	23	55
OPLS-AA	TIP4P-Ew[d]	0.57	2.25 ± 0.06	0.14	0.27	163	95
OPLS_2005	TIP4P-Ew	0.55	1.96 ± 0.04	0.14	0.28	150	97
AMBER ff99SB-ILDN	TIP3P	0.45	1.04 ± 0.01	0.06	0.16	28	34
AMBER ff03	TIP3P	0.36	0.65 ± 0.00	0.04	0.12	13	24
CHARMM27	TIP3P	0.44	1.00 ± 0.01	0.07	0.18	53	63
CHARMM22*	TIP3P	0.34	0.59 ± 0.00	0.04	0.13	28	55
OPLS-AA	TIP3P[d]	0.55	3.85 ± 0.12	0.27	0.55	278	105
OPLS_2005	TIP3P	0.55	3.37 ± 0.10	0.26	0.52	255	105
CHARMM27	mTIP3P	0.39	0.77 ± 0.01	0.05	0.14	49	74
CHARMM22*	mTIP3P	0.29	0.48 ± 0.00	0.03	0.11	26	62
OPLS_2005	TIP4P	0.57	3.07 ± 0.07	0.21	0.42	145	63
AMBER ff03	TIP4P/2005	0.29	0.46 ± 0.00	0.02	0.08	12	31
AMBER ff99SB-ILDN	SPC/E	0.42	0.84 ± 0.01	0.04	0.10	28	41
CHARMM27	SPC/E	0.40	0.78 ± 0.01	0.04	0.11	46	69
CHARMM22*	SPC/E	0.31	0.52 ± 0.00	0.02	0.09	27	59
OPLS_2005	SPC/E	0.58	2.13 ± 0.04	0.11	0.22	196	120

[c] Standard errors of P_{bound} for the 1:1 and 1:2 complex and K_A for the 1:2 complex were uniformly ≤ 0.01, while those of the mean lifetimes of the bound and unbound states were ≤ 3 ps.

[d] Simulations were run with the Desmond dynamics engine rather than the Anton dynamics engine, using the particle mesh Ewald method[61] rather than the Gaussian split Ewald method[63] for treatment of long-range electrostatics. The two dynamics engines and simulation protocols yield the same results, as verified by simulations of guanidinium acetate with the CHARMM22* force field and mTIP3P water model, which yield both K_A and P_{bound} values that agree to within the computed standard error.

[e] Experimental K_A values of guanidinium and butylammonium acetate permit only a qualitative estimate of the associated error. Taking two experimentally measured K_A values of guanidinium acetate using different protocols into account,[51,69] we estimate an error of ± 0.05, although the true uncertainty is not known. Using this estimate, we have

Table 2.4. Association constants (K_A) and probabilities of binding (P_{bound}) for guanidinium acetate from simulations involving a single molecule of guanidinium and acetate (0.1 M concentrations) without any chloride ions, using six biomolecular force fields with the TIP4P-Ew explicit water model. Also shown are the K_A and P_{bound} values computed from simulations containing 0.02 M acetate, 0.9 M guanidinium, and 0.9 M chloride ion. Results are from 1-μs simulations and standard errors were calculated using a block averaging method.[64]

Force Field	Water Model	0.1 M Guanidinium 0.1 M Acetate No Chloride Ions[f]		0.9 M Guanidinium 0.02 M Acetate 0.9 M Chloride Ion	
		P_{bound}[g]	K_A (M^{-1})	P_{bound}[g]	K_A (M^{-1})
AMBER ff99SB-ILDN	TIP4P-Ew	0.20	2.91 ± 0.12	0.57	2.23 ± 0.03
AMBER ff03	TIP4P-Ew	0.11	1.26 ± 0.04	0.45	1.12 ± 0.01
AMBER ff13α	TIP4P-Ew	0.19	2.74 ± 0.10	0.54	2.28 ± 0.03
CHARMM27	TIP4P-Ew	0.31	6.01 ± 0.32	0.61	4.06 ± 0.10
CHARMM22*	TIP4P-Ew	0.14	1.69 ± 0.06	0.48	1.31 ± 0.02
OPLS_2005	TIP4P-Ew	0.34 ± 0.02	7.23 ± 0.50	0.63	4.92 ± 0.17
Experiment[51,69,][h]		~0.04 ± 0.01	~0.37 ± 0.05	~0.25 ± 0.03	~0.37 ± 0.05

Table 2.5. Dielectric constants (ε_{water}) of pure water solutions simulated using the same protocol as for the amino acid analogue systems. Simulations were 100-ns in duration. Standard errors were calculated using a block averaging method.[64]

Water Model	ε_{water}
TIP4P-Ew	56.1 ± 0.4
TIP3P	102.1 ± 0.6
mTIP3P	109.4 ± 0.7
TIP4P	52.2 ± 0.3
TIP4P/2005	67.0 ± 0.5
SPC/E	69.7 ± 0.5
Experiment[74]	78.4 ± 0.1

back-calculated the range of simulated P_{bound} values that would be expected in our simulation, based on the experimental K_A.

[f] Simulations were run with the Desmond dynamics engine rather than the Anton dynamics engine, using the particle mesh Ewald method[61] rather than the Gaussian split Ewald method[63] for treatment of long-range electrostatics.

[g] Standard errors of P_{bound} were uniformly ≤ 0.01, unless otherwise noted.

[h] Experimental K_A values of guanidinium and butylammonium acetate permit only a qualitative estimate of the associated error. Taking two experimentally measured K_A values of guanidinium acetate using different protocols into account,[51,69] we estimate an error of ± 0.05, although the true uncertainty is not known. Using this estimate, we have back-calculated the range of simulated P_{bound} values that would be expected in our simulation, based on the experimental K_A.

2.7 SUBSEQUENT DEVELOPMENTS

Based on the promising results I obtained with the CHARMM22* and AMBER ff13α force fields, I decided to focus my production simulation of MoCVNH3 on these two force fields. Since ff13α was still a work in progress at the time, I reached out to the force field's developers about the availability of the final version, called ff14ipq. Upon receipt of the final force field, I first tested where or not our conclusions about ff13α applied to ff14ipq by simulating association of the three oppositely charged side-chain analogue pairs, as well as the Arg and Asp dipeptides. I discovered, as described in the next chapter, that changes made late in the development of ff14ipq led to severe overstabilization of salt bridge interactions. The developers asked us to collaborate on the development of an updated version, eventually named ff15ipq, leading to the work described below.

3.0 THE AMBER FF15IPQ FORCE FIELD FOR PEPTIDES AND PROTEINS

This chapter is based on a research article previously published as: Debiec, K. T.; Cerutti, D.S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T. Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model. *Journal of Chemical Theory and Computation*, 2016, 12, 3926-3947.

3.1 INTRODUCTION

Akin to developments spurred by the rapid expansion of computer power around 2000, the burgeoning capacity provided by programmable graphics processing units (GPUs) has extended the utility of molecular simulations as a practical tool for assessing biophysical processes.[78,79] Notably, GPU-accelerated computing has enabled routine simulations on the microsecond (μs) time scale, a critical regime on which biological processes including protein recognition, ligand binding, and protein conformational changes occur.[80] Access to these longer time scales may reveal flaws in the simulation models that were not previously apparent, driving refinements and leading toward improved predictive power of these models.

Historically, efforts in force field development have been largely focused on selecting only a subset of the parameters in a complex model for reoptimization, e.g., reoptimizing certain torsion parameters while keeping the set of atomic charges fixed to improve the accuracy in modeling particular behaviors, while retaining what is already successful. However, such efforts are limited by the accuracy of the unoptimized parameters. Many recent force field updates have focused on

refinements of torsion parameters.[28,32,81–89] However, these refinements may be compensating for deficiencies in the modeling of electrostatic and nonbonded interactions that limit the maximum attainable accuracy of the model. In addition, contemporary force fields have a tendency to borrow from a similar set of values for bond lengths, angles, and atomic radii that were fit many years ago, leading to interdependencies that may be difficult to untangle when optimizing only a portion of the parameters.

More recently, semiautomated schemes have been developed to simultaneously optimize hundreds of parameters thereby enabling the rapid development of new force fields. In particular, the Force Balance[90] and Implicitly Polarized Charge (IPolQ) methods[33,91] have yielded the AMBER ff15fb[92] and ff14ipq force fields,[91] respectively. These methods rely largely on automated tools for parameter optimization, but still require some amount of manual intervention in the form of fitting set composition or user-specified settings of the fitting algorithm. The engine behind the IPolQ workflow is the mdgx module of the AMBER software package.[93] This module combines its molecular dynamics (MD) facility with linear algebra routines for solving least-squares problems, manages extensive bookkeeping to organize parameters, provides user control over the fitting process, and interprets statistics to aid in further refinements. The mdgx module contains charge and torsion fitting routines that were built throughout the development of AMBER ff14ipq, the first complete protein force field based on the IPolQ scheme.[91]

Here, we have developed the new AMBER ff15ipq force field using the IPolQ workflow. The original motivation for the development of ff15ipq was to tackle concerns that its predecessor, ff14ipq, overestimates the stability of salt bridge interactions – a limitation shared with many other contemporary force fields.[56] However, in contrast to recently developed variants of force fields that address such concerns,[32,94,95] ff15ipq is far from a limited adjustment of its predecessor.

Rather, ff15ipq is a complete rederivation, comprising new atomic charges, a greatly expanded torsion parameter set with several new atom types to decouple distinct amino acids, and new backbone angle bending terms. In addition, whereas ff14ipq employed the TIP4P-Ew model for the solvent in the IPolQ scheme,[27] ff15ipq uses SPC/E_b, a recently developed three-point water model that yields more accurate rotational diffusion for proteins in solution.[96] The use of a three-point water model instead of a four-point water model, leaving out a virtual site, affords a modest improvement in the speed of CPU-based simulations and a larger acceleration in computing under the AMBER GPU engine.[97] In addition, the more-accurate rotational diffusion afforded by SPC/E_b opens new avenues for validating ff15ipq through direct calculation of NMR relaxation parameters. With the aid of GPUs and the AMBER GPU engine, we have extensively validated the force field by running MD simulations of peptide and protein systems on the μs time scale, yielding over 200 μs of aggregate simulation time.

We expect ff15ipq, or a close relative, to be valuable for a long time, even as we explore more expensive alternatives by adding virtual sites to both the protein and the standard water model. In addition, we are working to apply the mdgx workflow to other classes of biopolymers such as carbohydrates and nucleic acids, and to small organic molecules. We hope that the sweeping reoptimization made possible by mdgx and tools similar to it will inspire initiatives with other force fields and create complete chemical representations with predictive power in biomolecular simulations. Of future interest will be comparisons to contemporary force fields that have been developed in the traditional manner such as AMBER ff14SB,[81] OPLS-AA/M,[88] and CHARMM36,[87] as well as those developed using alternative sweeping reoptimization schemes, such as AMBER ff15fb.[92]

3.2 THEORY

3.2.1 The IPolQ method of force field parametrization

The Implicitly-Polarized Charge (IPolQ) method is a protocol for parametrizing fixed-charge force fields for solution-phase simulations that is comprised of two main components, implemented in the mdgx program of AmberTools.[93] The first component is a protocol for deriving nonpolarizable atomic charges that implicitly represent the energy of polarization by the presence of a solvent such as water.[33] The IPolQ charge derivation draws on approximations of dipole interactions in an external electrostatic field to arrive at the optimal nonpolarizable representation of a solute's atomic charges in the presence of a solvent such as water: precisely halfway between the charges that would reproduce the solute's electrostatic field in the gas phase and those that would reproduce the solute's electrostatic field after solvent-induced polarization.[98] This averaging comes about from the fact that the energy of a set of polarizable dipoles in an external field is identical to the energy of a set of fixed dipoles whose polarizations are halfway between the field-polarized dipoles and their gas-phase counterparts. IPolQ fits such fixed charges by applying the Restrained Electrostatic Potential (REsP) method,[99] using a pair of representations of the solute's electrostatic field corresponding to the vacuum and solution phases. While the former representation is straightforward to obtain from QM calculations, the latter is computationally unfeasible using a pure QM representation. Instead, the IPolQ method represents the polarizing Solvent Reaction Field Potential (SRFP) in its QM calculation using a field of point charges, derived from an MD simulation in which water, represented by the model with which the solute will ultimately be simulated, moves in equilibrium around the fixed solute.[33] Atomic charges are subsequently fit to reproduce the QM electrostatic potential at a set of grid points surrounding the molecule. As

described previously,[33] grid points are selected within the first and second solvation shells, with the inner boundary defined by excluding points for which the energy of the Lennard-Jones interaction between the solute and a probe representing the water model exceeds a selected maximum cutoff. While, in this work, we have applied equal weights to all of the selected grid points, others have found that more consistent charges may be obtained by applying a weighting function to de-emphasize points close to or distant from the solute.[100] Such improvements will be investigated as the IPolQ method is applied to other classes of molecules beyond peptides and proteins.

The second component of the IPolQ method is an extension for the fitting of bonded parameters that accounts for the discrepancy between the desired solution-phase conformational preferences and vacuum-phase QM calculations. This is accomplished by fitting a pair of solute charge sets: one appropriate for the vacuum phase (Q_{vac}), and the other for the solution phase (Q_{solv}). In the presence of the Q_{vac} charge set, the force field's bonded parameters are fit to reproduce the relative vacuum-phase QM energies of a diverse set of solute conformations. In subsequent simulations in the solution-phase, these same bonded parameters are paired with the polarized Q_{solv} charge set with the intention that the difference in the charge sets would account for the difference in solute conformational preferences between the vacuum and solution phases.[91]

3.2.2 Choice of water model for rederivation of IPolQ atomic charges

In contrast to the standard REsP method of fitting atomic charges for AMBER force fields,[99] the IPolQ method explicitly considers the influence of the water model on the solute's charge distribution.[33] While the atomic charges of the ff14ipq force field were fit using the TIP4P-Ew water model,[27] we have elected to fit the charges of ff15ipq using the SPC/E_b water model.[96] This

recently developed water model offers two advantages: in addition to the reduction in computational cost that is obtained by switching from a four-point to a three-point water model, this water model has been parametrized to yield accurate rotational diffusion of solvated proteins.

A key advantage to carrying out simulations with accurate rotational diffusion is the ability to directly calculate the NMR relaxation parameters ^{15}N R_1 and R_2 and ^{15}N-^{1}H heteronuclear NOE. These parameters provide information about fast dynamics (picosecond(ps), or nanosecond (ns) scale) of individual backbone N-H bond vectors within a protein, potentially offering a powerful means with which to validate MD simulations.[101] In principle, these NMR relaxation parameters may be calculated directly from an MD trajectory from the autocorrelation functions of the backbone N-H vectors. In practice, however, the poor reproduction of protein rotational diffusion in MD simulations using popular water models such as TIP3P limits the utility of such calculations.[29,96] This limitation has historically been addressed using approaches such as model-free analysis that attempt to separate the global rotational diffusion of the protein from its residue-specific internal dynamics, comparing only the internal dynamics between experiment and simulation. However, these approaches require extensive fitting of models to the experimental data, and further require that the global and local dynamics occur on separable time scales, limiting their applicability to highly flexible systems such as disordered peptides and proteins. Therefore, it would be preferable for MD simulations to yield accurate rotational diffusion, such that the simulated and experimental relaxation parameters can be compared directly.

Our decision to fit the charges of ff15ipq to the SPC/E_b water model was based on preliminary tests in which we ran a series of 24 simulations of the proteins GB3, ubiquitin, and binase using two different force fields (AMBER ff99SB-ILDN and CHARMM22*)[28,32] paired with four different water models (TIP3P, TIP4P-Ew, TIP4P-D, and SPC/E_b)[27,70,96,102] (Figure

3.13). Consistent with prior published work,[29,96] TIP3P and TIP4P-Ew yielded rotational diffusion significantly faster than experiment, and SPC/E_b yielded the most accurate result.

3.2.3 Extensions supporting restrained angle fitting

Earlier versions of mdgx were capable of fitting angle stiffnesses alongside torsions, but recent advances proposed by Vanommeslaeghe *et al.* permit the calculation of both the optimal stiffness constant and equilibrium value from the same linear least-squares problem.[103] The strategy generalizes work by Hopkins and Roitberg,[104] representing the parabolic angle as the sum of two parabolic basis functions and solving for both scaling coefficients to interpolate the optimal parameters. Basis functions were chosen such that their minima lay at ±0.2 radians from the equilibria of the original angles, which had been inherited from the AMBER ff94 force field.[105] As explained by Vanommeslaeghe *et al.*, the optimized angle's stiffness is given by the sum of coefficients solved for each basis function, and its equilibrium is given by the average of the two basis functions' minima weighted by their coefficients. Restraints on the optimized angle parameters follow from these definitions: a restraint equation setting the sum of the two coefficients $C_{i,1}$ and $C_{i,2}$ to a target value K_i, such as the stiffness of the original angle in the input force field, will harmonically penalize solutions which depart from the original stiffness value:

$$\alpha_{scl} N_i \alpha_{cpl} \left[C_{i,1} + C_{i,2} \right] = \alpha_{scl} N_i \alpha_{cpl} K_i \tag{3.1}$$

A similar restraint on the ratio of the two coefficients can be used to penalize solutions which depart from the original equilibrium value T_i, if the minima of the basis functions scaled by $Ci_{,1}$ and $Ci_{,2}$ are $B_{i,1}$ and $Bi_{,2}$, respectively:

$$\alpha_{scl} N_i \left[C_{i,1} \left(T_i - B_{i,1} \right) + C_{i,2} \left(T_i - B_{i,2} \right) \right] = 0.0 \tag{3.2}$$

As explained in the previous study,[91] these restraint equations will have a more pronounced effect if the data set contains only 10 rather than 1000 data points. Therefore, both sides of each restraint equation were scaled by a user-defined constant (α_{scl}) times the number of instances in which each optimizable angle appeared in the data set N_i, analogous to the scaling constant applied to torsion restraints. The scaling constants may appear to have no effect on the solution to the equations when either these constants are present on both sides of the equation or one side of the equation is zero. However, since the least squares fit finds an approximate solution to each equation, the scaling constants do, in fact, influence the relative importance of each restraint. In addition, due to the fact that these restraints penalize numerical deviations from the target values but angle stiffnesses and equilibria are expressed in different units by numbers of different scale, a separate scaling factor (α_{cpl}) was introduced to control the way in which restraints on the equilibria scale relative to restraints on the stiffness. For example, an α_{cpl} of ~57 applied to restraints on equilibria would penalize 1° deviations from the original value by the same amount as a 1 kcal/(mol rad^2) deviation from the original stiffness constant. After some experimentation, however, we found that, in our very large and heterogeneous data sets, much smaller values of α_{cpl} (0.5 to 1.0) result in the best fits to the data while partitioning the changes between equilibria and stiffness constants.

3.2.4 Addition of new atom types

Alongside the fitting of torsions and angles, several new atom types were added to ff15ipq in order to more accurately capture residue-specific conformational preferences. Most protein force fields use Gly and Ala as templates to develop backbone torsion parameters that are then inherited by other residues. The AMBER IPolQ force fields adopt an unconventional, concerted approach in which Φ, Ψ, Φ', Ψ', and all other torsions are simultaneously fit to the conformational preferences

of all residues in which they appear. While the resulting backbone torsions therefore consider the conformational preferences of residues other than Gly and Ala, their overall accuracy may decrease as different residues pull the parameters in different directions. Within the context of the IPolQ fitting method, a set of backbone torsion parameters that more accurately capture the conformational preferences of different residues may be obtained by introducing new atom types, creating decoupled classes of backbone torsions, which are applied to subsets of residues.

During the development of ff14ipq, three such classes were introduced: one for Gly, one for Pro, and one for all other residues. In order to decouple the Φ and Ψ torsions of Gly, which lacks Cβ and therefore has no Φ' and Ψ' torsions, a unique Cα atom type was assigned.[91] Similarly, the backbone torsions of Pro were decoupled by assigning a unique atom type for the backbone N. This additional atom type not only created unique Φ, Ψ, and Ψ' terms for Pro, but also a set of separate Ψ and Ψ' torsions for residues preceding Pro, thereby enabling the force field to capture the unique conformational preferences of these contexts.[106] The remaining residues were further divided into three subclasses based on their Cβ types, which determine the applied Φ' and Ψ' torsions. The Cβ types of ff14ipq yielded four subclasses: (i) flexible positively charged residues (Arg, Lys), (ii) residues whose Cβ atoms are bonded to two heavy atoms and whose side-chains are not aromatic (Asn, Asp, Cys, Gln, Glu, Leu, Met, Ser), (iii) residues whose Cβ atoms are bonded to three heavy atoms (Ile, Thr, Val), and (iv) all other residues (Ala, His, Phe, Trp, Tyr). Notably, this last subclass yielded Φ' and Ψ' torsions shared between Ala and the bulky aromatic residues. This unusual coupling was a consequence of the force field's lineage from ff12SB, where refitting of X_1 torsions alongside fixed Φ, Ψ, Φ', and Ψ' did not require a unique Cβ type for the aromatic residues, which already have unique Cγ types that yield unique X_1.[81]

In order to further improve the accuracy of residue-specific conformational preferences in ff15ipq, several new atom types were added to further decouple the backbone torsion parameters of different residues, leading to a total of five backbone classes. In order to restrict each class of backbone torsions to a single set of scaled 1-4 electrostatic terms, negatively-charged (Asp, Glu) and positively-charged (Arg, Lys) residues have been given unique Cα types, decoupling their Φ, Ψ, Φ', and Ψ' from those of the neutral residues. While the backbone N of Pro was decoupled in ff14ipq, it retained a shared Ψ' torsion; to break this dependency, Pro has now been assigned a new Cα atom type. Finally, the coupling between Ala and the bulky aromatic residues has been removed by assigning His, Phe, Trp, and Tyr a unique Cβ type, decoupling their Φ' and Ψ' terms from Ala. This decoupling divides the neutral residues into four subclasses: (i) Ala (ii) residues whose Cβ atoms are bonded to two heavy atoms (Asn, Cys, Gln, Leu, Met, Ser), (iii) residues whose Cβ atoms are bonded to three heavy atoms (Ile, Thr, Val), and (iv) bulky aromatic residues (His, Phe, Trp, Tyr). The backbone torsion classes of the 28 residue forms supported by ff15ipq are listed in Table 3.4.

3.3 METHODS

3.3.1 Calculation of the probability of binding (P_{bound}) for salt bridges

To compare the accuracy of ff15ipq in modeling the stability of protein salt bridges to its predecessor ff14ipq and contemporary force fields, we simulated the association of three pairs of oppositely charged amino acid side-chain analogues: guanidinium cation/acetate anion (Arg/Asp), butylammonium cation/acetate anion (Lys/Asp), and imidazolium cation/acetate anion

(His(+)/Asp). For comparison, such simulations were carried out using the polarizable force fields CHARMM Drude-2013 and AMOEBA,[107,108] in addition to other fixed-charge force fields that we had previously tested using these three model systems, as described in Chapter 2.[56]

Simulations with ff15ipq, ff14ipq, and AMOEBA were carried out using the AMBER 15 software package,[93] while those with CHARMM Drude-2013 were run with NAMD 2.10.0,[109,110] following a protocol analogous to that used for the previously evaluated fixed-charge force fields (full details are provided in Section 3.7.1).[56] Systems were constructed to be consistent with the experimental conditions under which the association constants (K_A) of guanidinium acetate and butylammonium acetate have been measured,[51] *i.e.,* each system consisted of 100 molecules of cation (guanidinium, butylammonium, or imidazolium), 2 molecules of acetate, and 98 chloride counterions solvated by ~18,000 water molecules. For the fixed-charge force fields, parameters of the side-chain analogues were based on those of the complete amino acids. For the CHARMM Drude-2013 polarizable force field, parameters of guanidinium, imidazolium, and acetate were those distributed alongside the force field.[107] Since methylammonium rather than butylammonium was used as the analogue of Lys during the development of Drude-2013,[107] the butylammonium acetate system was not tested with this force field. For the AMOEBA force field, parameters of guanidinium, imidazolium, and acetate were generated using the Poltype derivation protocol (details are provided in section 3.7.1).[111] As it was with CHARMM Drude-2013, the butylammonium acetate system was not tested with AMOEBA.

For all of the simulations mentioned above, the probability that an acetate molecule was bound to one or more cation molecules was calculated by assigning each pair to either the bound or the unbound state. For each force field and pair of side-chain analogues, definitions of the unbound and bound states were based on the potential of mean force (PMF) as a function of the

minimum distance between nitrogen atom(s) of the cation and the oxygen atoms of acetate. In particular, the cutoff between the bound and unbound states was defined as the point of inflection between the free energy minimum of the bound state (~2.5−3 Å) and the free energy maximum, which corresponds to the desolvation barrier (~3−3.5 Å). Pairs whose minimum N-O distances were below this cutoff were assigned to the bound state, while those beyond were assigned to the unbound state. In addition to species in which a single acetate molecule was bound to a single cation molecule, forming a 1:1 complex (*e.g.*, the guanidinium/acetate complex), species in which acetate was bound to two or more cation molecules (*e.g.*, the 2:1 diguanidinium/acetate complex) were observed and counted separately. Standard errors were calculated using a block averaging method.[64]

3.3.2 Rederivation of IPolQ atomic charges with the SPC/E_b water model

The atomic charges of ff15ipq were fit using the IPolQ module of mdgx as described previously for ff14ipq.[91] During charge fitting, each amino acid was represented by a blocked dipeptide including acetyl (Ace) and N-methylamide (Nme) caps; terminal forms were represented by omitting one of the blocking groups, while the disulphide form of cysteine (Cyx) was represented by a pair of dipeptides linked by a disulfide bond. To expand on the set of amino acids and protonation states that were supported by ff15ipq, atomic charges were also derived for the following: the N- and C-terminal forms of protonated aspartate (Ash) and glutamate (Glh), the C-terminal form of neutral lysine (Lyn), the terminal and nonterminal forms of deprotonated cysteine (Cym), and the noncanonical amino acid norleucine (Nle).

Each solute of interest was solvated in a cubic box of SPC/E_b water with a clearance of 10 Å between the solute and the edge of the box, and subjected to a high-temperature MD simulation at 450 K, from which were collected a set of 20 conformations. Each conformation was subsequently re-equilibrated at 298 K before being input to the IPolQ module of mdgx. This module was used to run an MD simulation with the solute fixed, during which the coordinates of surrounding solvent molecules were collected and used to generate a collection of point charges representing the solvent reaction field potential. This collection consists of an inner cloud of point charges taken directly from the coordinates of solvent molecules within 5 Å of the solute, and three outer shells of point charges fit to reproduce contributions to the solvent reaction field potential from the infinite periodic system beyond 5 Å.

A pair of QM calculations for the solute were then run at the MP2/cc-pVTZ level of theory:[112–115] one in vacuum and the other including the solvent reaction field potential as modeled by the collection of point charges. These calculations were run using the ORCA 3.0.3 software package for each conformation of each residue,[116] requiring over 3,000 density calculations. The resulting densities were then input to mdgx's FitQ module, yielding a pair of charge sets, one valid for simulation in vacuum (Q_{vac}) and the other for simulation in solution (Q_{solv}).

3.3.3 Generation and extension of the angle and torsion fitting dataset

The bonded parameters of ff15ipq were fit to reproduce the relative vacuum-phase QM MP2/cc-pVTZ potential energies of a set of diverse conformations of short peptides using an iterative cycle of refinement, similar to that used for its predecessor, ff14ipq.[91] This cycle involved the following steps: (i) MD simulations were carried out to generate a set of peptide conformations, (ii) these conformations were subjected to energy minimization in vacuum using the molecular mechanics

(MM) energy function with Q_{vac} and the current generation of bonded parameters, (iii) QM energies of the energy-minimized conformations were calculated, (iv) the conformations and energies were used to fit an improved set of bonded parameters, and (v) steps (i) through (iv) were repeated to fit the next generation of bonded parameters. In this way, subsequent generations of the force field "learned" from the biases of their ancestors, provided those biases were captured in the QM energies of the additional conformations that resulted from step (i) of the iterative cycle.

During the development of ff14ipq, selected conformations from an initial fitting set of ~28,000 were subjected to energy minimization with each new generation of bonded parameters to yield new conformations, accumulating a total of 65,000 structures and single-point energies.[91] The first generation of ff15ipq fitting data was created by pairing ff14ipq with generalized Born implicit solvent MD simulations[117] of amino-acid dipeptides at 450 K, followed by vacuum energy minimization of many snapshots from each simulation. While we have not tested how well ff14ipq behaves with implicit solvent, the purpose was to capture any spurious conformational preferences that might remain in the original force field. Additionally, we included ~1,400 conformations of the Ace-Ala-Pro-Ala-Nme tetrapeptide, while the second generation added numerous tripeptides containing Gly, and conformations of the disulfide-bridged Cys•Cys system (among the largest of all the systems used in QM single point energy calculations). These refinements added ~15,000 new conformations to the ff15ipq fitting set.

The next three generations of refinement were designed to cover sampling of the multiple classes of backbone parameters applied to different residues, as described in Section 3.2.4. In order to ensure sampling of diverse backbone conformations, conformations were generated by progressively restraining Φ and Ψ at 20° intervals using a 16 kcal/mol·rad^2 harmonic restraint over the course of the MD simulation, yielding 324 conformations of each. Since the unique backbone

nitrogen type of Pro creates unique Ψ and Ψ' terms for preceding residues, the third generation of conformations consisted of 51 Pro-containing tripeptides for which the non-Pro residue's Φ and Ψ were restrained. At this point in development, it was decided to branch the positively and negatively charged residues into unique backbone classes, and as such the fourth and fifth generations of refinement consisted of 57 tripeptides containing the charged residues Asp, Glu, Cym (deprotonated Cys), Arg, Lys, and Hip (doubly-protonated His), in which Φ and Ψ of the charged residues were restrained. In order to cover the unique backbone parameters applied to the terminal forms of each residue, additional conformations were added for a set of 78 terminal NXaa-Nme and Ace-CXaa monopeptides. For these terminal systems, scans of either the unique Ψ of the N-terminal forms or the unique Φ of the C-terminal forms were performed at 2° intervals, yielding 180 conformations of each. Since the unique backbone nitrogen types of the N-termini and Pro in tandem yield an additional set of Ψ terms for NXaa-Pro, scans of Ψ were run for an additional set of NXaa-Pro-Nme dipeptides. Finally, in order to cover the unique backbone Ψ and Ψ' terms of the amide blocking group (Nhe), scans of Φ and Ψ were run for 17 Ace-Xaa-Nhe dipeptides, yielding 324 conformations of each. During these three generations of refinement, ~60,0000 conformations were added to the ff15ipq fitting set.

After the fifth generation of refinement, support for the fitting of angle equilibria and force constants alongside torsions was implemented in mdgx, and subsequent generations emphasized comprehensive sampling of backbone angles. The sixth, seventh, and eighth generations of refinement consisted of perturbations of the angles around N, Cα, and C. Starting from an initial conformation, a selected angle of interest was subjected to a random perturbation within a range of ±20° of its original equilibrium value (as inherited from the ff94 force field and retained in contemporaries such as ff14SB). Target values for the other angles around the same central atom

were then chosen by taking their initial values and adjusting them such that the total sum of angles around the central atom was appropriate for the known geometry; target sums of 360° for planar geometry around N and C and 660° for tetrahedral geometry around Cα were used. During subsequent MM minimization, the target values for these angles were restrained using 256 kcal/mol•rad^2 harmonic restraints. During the eighth and final generation of refinement, angle perturbations were resampled in the context of new scans of Φ and Ψ backbone torsions at 10° intervals for each Ace-Xaa-Nme dipeptide, yielding 1296 conformations of each, alongside additional sampling of terminal monopeptides. During these three generations, ~125,000 conformations were added, yielding a final fitting set of ff15ipq consisting of >250,000 single-point QM energies, which is over four times larger than that used for ff14ipq.

3.3.4 Fitting of torsion and angle terms

As done previously for ff14ipq, the torsion parameters of ff15ipq were fit using a linear least-squares fit implemented in the Param module of mdgx;[91] extensions to the module for angle fitting are described in Section 3.2.3. This module selects a set of torsional barrier heights, angle equilibria, and angle stiffnesses that best reproduce the relative conformational energies of the systems included in the fitting set. During the fitting process, the Fourier series lengths and phase angles of the torsional terms were not optimized, and phase angles were set to either 0° or 180° to enable the development of parameters that are transferable to alternative chiralities. All backbone Φ, Ψ, Φ', Ψ', and side-chain X torsions of nonterminal forms of the amino acid residues were allocated four terms in their Fourier series. Torsions unique to the terminal forms of residues and residues preceding Pro were restricted to only three terms since these terms were less exhaustively sampled in the fitting set. This restriction was applied to limit the risk of overfitting. While all

torsion parameters of residues in the fitting set were fit, only angles in which the central atom was the N, Cα, or C of a nonterminal residue were fit. For Pro, only the angles around Cα were fit since the unique backbone N type of Pro introduces a large number of parameters that depend on the preceding residue. In order to avoid overfitting torsional barrier heights, torsions were restrained towards 0° with a force constant of $2x10^{-4}$ kcal/mol. Similarly, angles were restrained to their original values, inherited from ff94, with the equilibria and stiffness force constants set to $5x10^{-5}$ kcal/mol and $2x10^{-4}$ kcal/mol, respectively.

3.3.5 Umbrella sampling of tetrapeptides

To characterize the backbone conformational preferences of ff15ipq in explicit SPC/E_b water, we carried out umbrella sampling simulations of blocked tetrapeptides Ace-Ala-Xaa-Ala-Nme, calculating the potential of mean force as a function of the backbone Φ and Ψ torsions of the central residue Xaa. In order to identify differences in the conformational preferences between ff15ipq, its predecessor ff14ipq, and contemporary force fields, simulations of Ace-Ala-Ala-Ala-Nme were carried out using the AMBER force fields ff15ipq, ff14ipq,[91] and ff14SB;[81] the OPLS force field OPLS-AA/M;[88] and the CHARMM force fields CHARMM36 and Drude-2013.[87,107] Analogous simulations were carried out to compare the conformational preferences of other central amino acid residues using the AMBER ff15ipq, ff14ipq, ff14SB, and CHARMM36 force fields. The backbone Φ and Ψ torsions of the central residue were restrained in a series of 1296 windows spaced at 10° intervals, using a harmonic penalty function with a force constant of 8 kcal/mol·rad^2. Each window was seeded from a continuous, incrementally restrained simulation, and sampled for 2.0 ns following a 0.2-ns equilibration. From each set of 1296 windows were reconstructed the unbiased potentials of mean force using the weighted histogram analysis method (WHAM).[118,119]

3.3.6 Simulations of benchmark systems

To validate ff15ipq as a general force field for peptides and proteins, extensive MD simulations on the μs time scale were carried out for a variety of benchmark systems consisting of both structured and disordered peptides and proteins. For each system, the amino-acid sequence or PDB code, sources of initial coordinates, and temperatures maintained throughout the simulations are listed in Table 3.1. Further details of the benchmark systems are provided below.

Table 3.1. Peptide and protein validation systems.

System	Sequence/PDB	Residues	Temperature (K)	Duration (μs)
Ala_5	$^{+}$AAAAA$^{\varnothing}$	5	298	6
K19	Ace-GGG-$(KAAAA)_3$-K-Nhe[120]	19	275, 285, …, 315, 325	4
$(AAQAA)_3$	Ace-$(AAQAA)_3$-Nhe[121]	15	280, 290, …, 320, 330	4
GB1 Hairpin	$^{+}$GEWTYDDATKTFTVTE^{-}[122]	16	275, 285, …, 315, 325	4
Chignolin	$^{+}$GYDPETGTWG^{-}, 1UAO[123]	10	298	4
Cln025	$^{+}$YYDPETGTWY^{-}, 2RVD[124]	10	280, 290, …, 360, 370	4
Trp-cage	1L2Y[125]	20	275, 285, …, 315, 325	4
Binase	1BUJ[126]	109	298	10
BPTI	5PTI[127]	58	298	10
GB3	1P7E[128]	56	298	10
Lysozyme	4LZT[129]	129	300	2
Ubiquitin	1UBQ[130]	76	298	10
Villin Headpiece[i]	2F4K[131]	35	303	10
P53[j]	1YCR[132]	13	298	10
P53/MDM2[b]	1YCR[132]	13/85	298	10
S-peptide[k]	1RNU[133]	22	298	10
S-peptide/S-protein[c]	1RNU[133]	22/104	298	10

[i] HP35 double-norleucine mutant mutant (Lys24Nle, Asn27His, and Lys29Nle)

[j] The p53 peptide used contained residues 17-29 of the full-length protein and included an N-terminal acetyl (Ace) and C-terminal amide (Nhe) blocking group. MDM2 included residues 25-109 of the full-length protein, omitting a mobile N-terminal region unresolved in the crystal structure. The N- and C-termini of MDM2 were blocked with acetyl (Ace) and N-methylamide (Nme) blocking groups, respectively.

[k] In order to accurately match the amino acid sequences used in NMR experiments,[215,216] residues not resolved in the crystal structure were built using Avogadro.[217] Residues STSAA were appended to the C-terminus of the S-peptide, and SSS to the N-terminus of the S-protein. Additionally, residues GA were appended to the N-terminus of the S-peptide, representing a cloning artifact present in the NMR experiments. These residues were not restrained during equilibration of the system. The NMR experiments on the S-peptide/S-protein complex were conducted at pH 3.7, making the appropriate protonation state of Asp and Glu, whose pK_as average ~3.7 and ~4.1, uncertain within the confines of our model which does not capture proton exchange. We therefore elected to run our simulations with Asp deprotonated and Glu deprotonated; His was protonated in all simulations.

3.3.6.1 Structured peptides and proteins

As done for ff14ipq,[91] we validated ff15ipq by simulating penta-alanine (Ala_5), the α-helical K19 peptide, the GB1 β-hairpin from the C-terminal fragment of Protein G, the designed β-hairpin chignolin, Trp-cage, GB3, and lysozyme. We also carried out simulations of the α-helical $(AAQAA)_3$ peptide, the Cln025 mutant of the chignolin β-hairpin, the double-norleucine variant of the villin headpeice subdomain, bovine pancreatic trypsin inhibitor (BPTI), ubiquitin, and binase.

3.3.6.2 Disordered peptides

In order to evaluate the ability of ff15ipq to model disordered proteins, we simulated two classic systems for studying the binding processes of disordered peptides that fold only upon binding their partner proteins: (a) the N-terminal p53 peptide and MDM2 oncoprotein, and (b) the S-peptide and S-protein cleavage products of the RNAse A protein. Both of these peptides fold into α-helical conformations only upon binding their partner proteins. Simulations were performed with these peptides both in isolation and in complex with their protein binding partners.

3.3.6.3 Simulation configuration and analysis

Simulations of benchmark systems were carried out using the GPU implementation of the *pmemd* module in the AMBER 15 software package.[97,134] Each system was solvated in a truncated octahedral box of SPC/E_b explicit water with a 12 Å buffer for the disordered peptide/protein systems and 10 Å buffer for all other systems. Prior to production simulation, each system was subjected to energy minimization followed by a three-stage equilibration. In the first stage, a 20-ps simulation of the energy-minimized system was carried out at constant temperature while restraining the solute heavy atoms to their initial positions using a harmonic potential with a force

constant of 1 kcal/(mol Å^2). In the second stage, a 1 ns simulation was carried out at constant pressure with the same harmonic position restraints. Finally, an additional 1 ns unrestrained simulation was carried out at constant temperature and pressure. Temperatures were maintained at selected values (between 270 and 370 K) using a Langevin thermostat (frictional constant of 1 ps^{-1}) while pressure was maintained at 1 atm using a Monte Carlo barostat (200 fs between attempts to change the system volume).[58] Van der Waals and short-range electrostatic interactions were truncated at 10 Å; long-range electrostatic interactions were calculated using the particle mesh Ewald method.[61] To enable at least a 2 fs time step, bonds to hydrogen were constrained to their equilibrium values using the SHAKE and SETTLE algorithms.[135,136] For the K19, $(AAQAA)_3$, GB1 hairpin, chignolin, and Cln025 systems, hydrogen mass repartitioning was used to enable the use of longer time steps.[137] In particular, the masses of solute hydrogen atoms were increased by a factor of three, and that of their attached heavy atoms decreased by a corresponding amount such that the total mass remained constant; the masses of water molecules were not repartitioned. This mass repartitioning scheme enables a 4 fs time step for simulations at ≤300 K; for simulations at >300 K, shorter time steps were used and set to be equal to 1200 K fs divided by the set temperature. Conformations were saved every ps for analysis by the AmberTools *cpptraj* program.[138] Diagnostics included DSSP,[139] rotational diffusion,[29] and NMR relaxation calculated by the iRED method.[140]

3.4 RESULTS

3.4.1 Strengths of protein salt bridges

To evaluate the accuracy of ff15ipq in modeling the strengths of protein salt bridges, we simulated the association of three pairs of oppositely charged amino acid side-chain analogues: guanidinium cation/acetate anion (Arg/Asp), butylammonium cation/acetate anion (Lys/Asp), and imidazolium cation/acetate anion (His(+)/Asp). For each salt bridge, the resulting probability of an anion binding to one or more cation molecules (P_{bound}) was compared to experiment (if available) as well as those of six other fixed-charge force fields, including ff14ipq, ff14SB, ff03, CHARMM22*, CHARMM36, and OPLS_2005, and two polarizable force fields, CHARMM Drude-2013 and AMOEBA.

The original motivation for the development of ff15ipq was to correct for the overstabilization of protein salt bridges by its predecessor, ff14ipq. During the development of ff14ipq, the Lennard-Jones radii of several polar heavy atoms were refit to reproduce the experimental solvation free energies of side-chain analogues.[33] Although the resulting set of radii were initially intended to be applied globally, several of the larger radii resulted in increased 1-4 repulsion during torsion fitting, which made the torsion parameters more difficult to fit. To overcome this difficulty, mixed Lennard-Jones combining rules (called LJEDIT within AMBER software or NBFIX within CHARMM) were applied to these polar groups, assigning different radii for their solute-solvent interactions from those used for their solute-solute interactions. For example, for the carboxylate oxygen atoms of the side-chains of Asp and Glu larger Lennard-Jones radii for interactions with water were used than for interactions with solute atoms.[91] An undesirable effect of this strategy, however, was the overstabilization of salt bridges – to the point that in our

simulations each acetate molecule was bound to three or more cation molecules (*e.g.*, guanidinium) for most of the simulation (Figure 3.1 and Figure 3.14). Essentially, the larger radii used for solute-solvent interactions forced the carboxylate group out of solution and into interactions with available solute atoms.

For ff15ipq, we addressed the problem of overstabilized salt bridges by discarding the mixed Lennard-Jones radii of ff14ipq and instead applied empirical corrections to the radii of polar hydrogen atoms bonded to nitrogen (atom type 'H') in both the protein backbone and side-chains (note that the original σ of 1.07 Å for this atom type may equivalently be expressed as an R* of 0.6000, and the details of its fitting appear to have been lost to history).[105,141] These corrections were determined from simulations of the three oppositely-charged side-chain analogue systems with H σ ranging from the ff94 value of 1.07 Å up to 1.5 Å, calculating the probability of salt bridge formation (P_{bound}), and comparing this probability to that from experiments. Based on the results (Figure 3.14) we selected a σ of 1.3 Å for nitrogen-attached hydrogens in both the protein backbone and side-chains. For guanidinium acetate, we found that a further increase in σ to 1.5 Å for the side chain of Arg was necessary to achieve satisfactory agreement with the experimental value of this system. All other Lennard-Jones radii retained their original ff94 values.

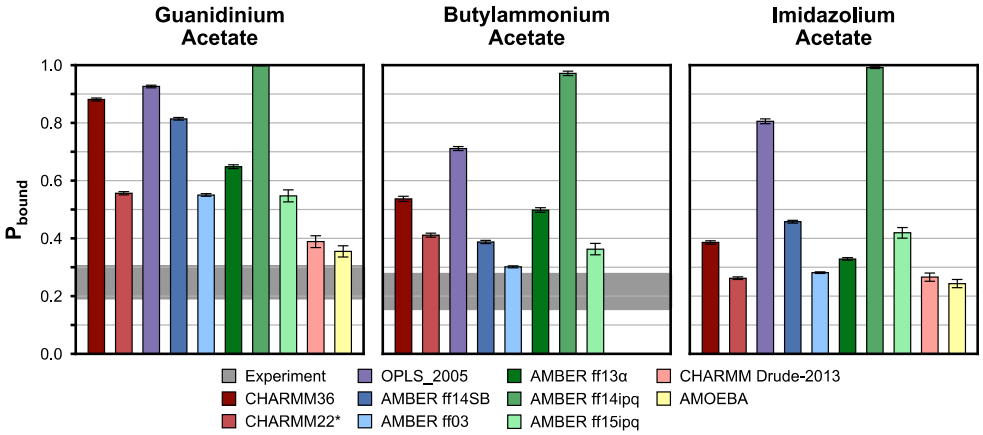

Figure 3.1. Probability of binding (P_{bound}) between acetate and one or more molecules of three cationic side-chain analogues using seven fixed-charge and two polarizable biomolecular force fields, each paired with either the water model with which it was derived or that with which it is most-commonly used. The P_{bound} values corresponding to the experimentally determined K_A values of guanidinium acetate and butylammonium acetate are depicted as horizontal gray bars;[51,69] no experimental value is available for the imidazolium acetate system. Error bars represent 95% confidence intervals calculated using a block averaging method.[64] Results for the CHARMM36, CHARMM22*, OPLS_2005, AMBER ff14SB, and AMBER ff03 force fields are from previous simulation studies described in Chapter 2.[56]

As shown in Figure 3.1, ff15ipq yields P_{bound} values that are among the most reasonable, relative to the other fixed-charge force fields that were tested (ff03, ff14SB, ff14ipq, OPLS_2005, CHARMM22*, and CHARMM36). For guanidinium acetate, our results with ff15ipq are roughly consistent with CHARMM22* and ff03, which we had previously found to provide the most accurate modeling of this system.[56] For butylammonium acetate, ff15ipq yields a P_{bound} that is slightly higher than that of AMBER ff03, but similar to those of CHARMM22* and ff14SB. For imidazolium acetate, ff15ipq yields a P_{bound} that is higher than those of CHARMM22* and ff03, but similar to that of ff14SB. Of particular note is that CHARMM22* was also parametrized to reproduce the experimental association of guanidinium acetate, but via adjustments to the atomic

charges of only the side-chains of Arg, Asp, and Glu;[32] as a result, these adjusted parameters are inconsistent with the rest of the force field, whose charges had been fit years earlier using a different method.[142] Along the same lines, a recently developed variant of the AMBER ff99SB-ILDN force field has involved the application of mixed Lennard-Jones combining rules exclusively to interactions between the side chains of Arg, Asp, and Glu.[94,95] In contrast to the post hoc adjustments of these two other force fields, our approach involves first adjusting the Lennard-Jones radii followed by refitting of atomic charges and bonded parameters. This approach – which has been an onerous one in the past – has been significantly streamlined by the mdgx software.

We note that all of the fixed-charge force fields are outperformed by the more expensive, polarizable CHARMM Drude-2013 and AMOEBA force fields. While it is likely that much of their superior performance results from the more complex charge model for the solutes, it is also possible that the solute-solvent interactions – which compete with solute-solute interactions – are more accurately represented by the use of polarizable water models. In particular, fixed-charge water models similar to the SPC/E_b model used here have recently been found to underestimate the strength of solute-solvent interactions in general,[102] and it is possible that this limitation of the water models restricts the accuracy with which the solute models may represent salt bridges.

3.4.2 Optimization of torsion and angle parameters

A key metric for assessing the accuracy of the torsion and angle parameters of ff15ipq was the ability to reproduce the target QM potential energy surface. Figure 3.2 shows the distribution and root mean square error (RMSE) of ff15ipq energies with respect to their target QM potential energies for the 20 canonical amino acids. The RMSE values for all neutral residues are <1.3 kcal/mol, while those of the negatively charged residues Asp and Glu are <1.9 kcal/mol, and those

of the positively charged residues Arg and Lys are <2.4 kcal/mol. As shown in Figure 3.15, the neutral forms Asp, Gln, and Lys (Ash, Glh, and Lyn, respectively) have RMSE values that are consistent with the other neutral residues, suggesting that the increased RMSE values of Lys and Arg, relative to uncharged residues, are related to their net charge, rather than to their additional flexible X torsions. Optimization of the backbone angle parameters introduced a 5-15% improvement in RMSE and enabled expansion of the fitting set by more than four-fold without sacrificing the ability to reproduce those parts of the QM potential energy surface represented in the original data set.

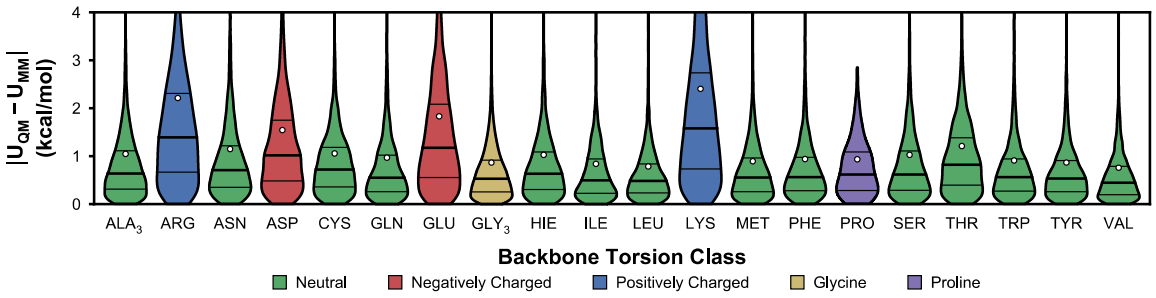

Figure 3.2. Distributions of residuals of relative molecular mechanical potential energies with respect to their quantum mechanical target potential energies for 18 Ace-Xaa-Nme dipeptides and the Ace-Ala-Ala-Ala-Nme and Ace-Gly-Gly-Gly-Nme tetrapeptides. 25th, 50th, and 75th percentiles are represented by horizontal lines, and means are represented by white circles. Each dataset is colored by its corresponding backbone torsion class (see **Table 3.4**).

3.4.3 Conformational preferences of individual residues and very short peptides

To assess the accuracy of ff15ipq in modeling the backbone conformational preferences of proteins within computationally tractable systems, we carried out a series of simulations of short peptides which may be affordably simulated to convergence. Initially, we focused on simulations of the

Ace-Ala-Ala-Ala-Nme tetrapeptide, for which we calculated the PMF as a function of the backbone Φ and Ψ torsions of the central residue. Figure 3.3 shows the results for ff15ipq, its predecessor ff14ipq, and several contemporary force fields. Relative to ff14ipq, ff15ipq has larger free energy barriers (by ~1 kcal/mol) between the α well (Φ ≈ -70°, Ψ ≈ -20°) and γ' well (Φ ≈ -80°, Ψ ≈ 60°) and between the β well (Φ ≈ -150°, Ψ ≈ 150°) and PPII well (Φ ≈ -70°, Ψ ≈ 140°). In addition, ff15ipq has a more clearly defined ξ well (Φ ≈ -140°, Ψ ≈ 50°). On the left half of the Ramachandran plot, the depth of the Lα well (Φ ≈ 60°, Ψ ≈ 40°) has decreased slightly, and that of the γ well (Φ ≈ 70°, Ψ ≈ -40°) has decreased by ~1 kcal/mol, while the PII' well (Φ ≈ 60°, Ψ ≈ -130°) has been retained. Relative to the ff14SB and CHARMM36 force fields, ff15ipq shows similar α and PPII well depths, though ff14SB and CHARMM36 do not exhibit γ' or ξ wells and the precise positions of the various wells differ between the force fields. Larger differences are observed relative to the OPLS-AA/M and polarizable CHARMM Drude-2013 force fields, which have shallower and deeper β wells, respectively.

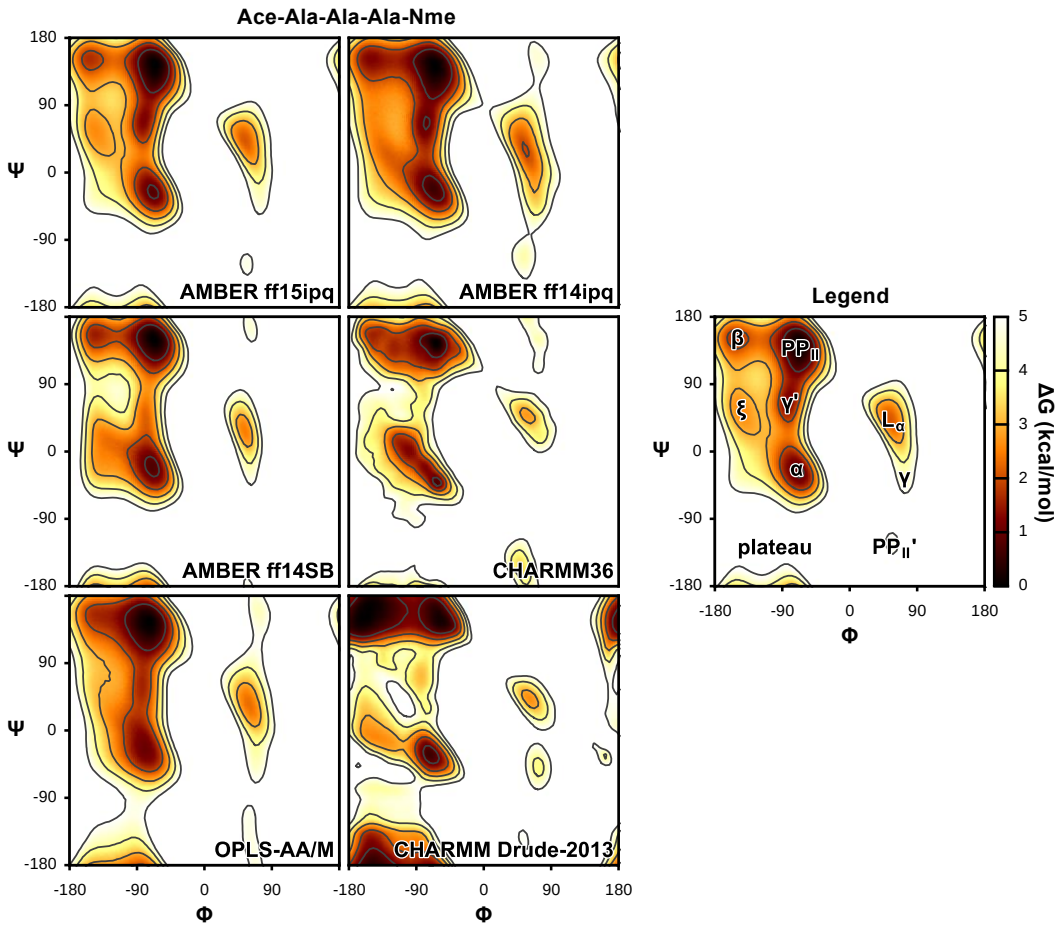

Figure 3.3. Potentials of mean force for the central residue of blocked alanine tetrapeptides as a function of backbone Φ and Ψ torsions of the central residue using five fixed-charge force fields and one polarizable force field. Each force field was paired with either the water model with which it was derived or that with which it is most-commonly used.

Next, we extended our validation of ff15ipq by examining residue-specific backbone conformational preferences. In particular, we carried out simulations of Ace-Ala-Xaa-Ala-Nme tetrapeptides containing each of the 20 canonical residues at the central position, including the 25 protonation states of these residues that are supported by the force field. For comparison, analogous simulations were carried out using the ff14ipq, ff14SB, and CHARMM36 force fields. The resulting Φ/Ψ backbone torsional preferences of the central residues were then compared to those of Ala-Xaa-Ala obtained from the Neighbor-Dependent Ramachandran Distribution

(NDRD) dataset, derived from conformations observed in the loops regions of proteins (non-α-helix/β-sheet secondary structures).[143] The NDRD dataset is drawn from a collection of ~3,000 high-resolution crystal structures in the Protein Data Bank,[144] and accounts for the influence of preceding and following residues on the Φ/Ψ backbone torsional preferences of the central residue. Given the considerable differences in the contexts of our simulations and the NDRD experimental dataset, *i.e.*, solution vs crystal environment, we focused solely on qualitative differences between the simulated and experimental conformational preferences of each peptide. In particular, we compared the conformational preferences of peptides containing a nonalanine central residue relative to that of the reference Ace-Ala-Ala-Ala-Nme peptide.

Generally, both ff15ipq and ff14ipq show greater variation between amino acids than ff14SB and CHARMM36, which apply the same backbone torsions to all residues (Figure 3.16). Several differences between ff15ipq and ff14ipq are apparent. For the neutral residues whose Cβ atoms are bound to two heavy atoms, the clearest difference is the decreased favorability of the $-180° < \Psi < -90°$ region for Asn, Gln, Leu, and Met; ff15ipq is more consistent with NDRD distributions in which such conformations are rare, due to the broader sampling of such uncommon backbone conformations in the ff15ipq fitting set. An exception is Ser, which retains this region and exhibits overall broader sampling, in contrast to the NDRD distribution, in which conformations are restricted largely to the canonical wells. For the neutral residues whose Cβ atoms are bound to three heavy atoms (Ile, Thr, and Val), the NDRD dataset shows increased conformational preferences in the β region and in the region adjacent to the α well, centered at $\Phi \approx -120°$, $\Psi \approx -60°$. These preferences are captured by both ff14ipq and ff15ipq, but the lower region is erroneously disfavored by both ff14SB and CHARMM36. While ff15ipq is improved compared to ff14ipq by disfavoring the Lα well of Thr, the near-absence of sampling of this well in the

NDRD distribution suggests that it may still be too favorable, relative to experiment. Differences between ff15ipq and ff14ipq for the bulky aromatic residues are less pronounced; the conformational preferences of these residues may be more dependent on sterics as modeled by the Lennard Jones parameters, which have not been changed in ff15ipq from those of ff14ipq. The greatest differences between ff15ipq and ff14ipq are observed for the negatively charged residues Asp and Glu, which have been granted their own Φ, Ψ, Φ', and Φ' torsions in ff15ipq. These residues largely restrict sampling to the of PPII, γ', and α wells, lacking clearly-defined β wells and any wells on the right side of the Ramachandran plot. The differences for the positively charged residues Arg and Lys are much smaller, likely because these residues were already assigned unique Φ' and Ψ' torsions in ff14ipq.

To complement the above qualitative comparisons, we obtained quantitative measures of the accuracy of ff15ipq's backbone conformational preferences by calculating J-coupling constants for the Ala_5 peptide and comparing these values to experiment. This peptide was the focus of a study by Best et al.[145] in which multiple force fields were compared in terms of their ability to reproduce experimental J-coupling constants using the Karplus equation and three different sets of Karplus coefficients: the original coefficients as used by Graf et al.[146–149] and two sets of DFT-based coefficients (DFT-1 and DFT-2) by Case et al.[150] In this study, a suggested criterion for a high-quality force field is that the χ^2 value between calculated and experimental J-coupling constants should be ≤2.25 for all three sets of Karplus coefficients. Three useful points of reference are the recently developed ff14SB, CHARMM36, and ff03w force fields, which were empirically corrected to improve reproduction of experimental Ala_5 J-coupling constants.[81,85,87] The ff14SB force field yielded χ^2 values of 0.9 and 1.2 with the original and DFT-2 coefficients,

respectively, but a higher χ^2 of 2.7 with the DFT-1 coefficients; CHARMM36 and ff03w were tested only with the DFT-2 coefficients, yielding χ^2 of 1.16 and 0.9, respectively.

As shown in Table 3.2, ff15ipq yields χ^2 of 0.53 with the original coefficients, χ^2 of 1.08 with the DFT-2 coefficients, and a χ^2 of 0.67 with an additional set of Karplus coefficients from Lindorff-Larsen et al.[151] However, like ff14SB, we obtained a higher χ^2 value of 2.91 with the DFT-1 coefficients; in our case, the higher χ^2 value is driven primarily by a single outlier that deviates greatly from experiment, $^3J_{HNC\beta}$. Based on results from preliminary versions of ff15ipq, it appears that lower χ^2 values with the original Karplus coefficients may come at the expense of higher χ^2 values with the DFT-1 coefficients. Notably, ff15ipq – which employs a general parametrization to reproduce QM potential energies – performs at least as well as ff14SB, CHARMM36, and ff03w, which have been parametrized specifically to reproduce Ala_5 J-coupling constants.[81,87] All four force fields yield results much improved relative to force fields developed only a few years ago.[145]

Table 3.2. Ala_5 J-coupling constants

		Simulation				
J-coupling	**Residue**	**Original**	**DFT-1**	**DFT-2**	**KLL**	**Experiment**
$^1J_{N,C\alpha}$	2	11.38	11.38	11.38	11.38	11.36
$^1J_{N,C\alpha}$	3	11.06	11.06	11.06	11.06	11.26
$^2J_{N,C\alpha}$	2	8.64	8.64	8.64	8.64	9.20
$^2J_{N,C\alpha}$	3	8.50	8.50	8.50	8.50	8.55
$^3J_{C,C}$	2	0.67	0.48	0.57	0.67	0.19
$^3J_{H\alpha,C}$	2	1.57	1.31	1.47	1.38	1.85
$^3J_{H\alpha,C}$	3	1.83	1.60	1.77	1.67	1.86
$^3J_{HN,C}$	2	1.26	1.27	0.88	1.46	1.10
$^3J_{HN,C}$	3	1.19	1.20	0.88	1.37	1.15
$^3J_{HN,C\beta}$	2	2.10	4.06	3.24	2.17	2.30
$^3J_{HN,C\beta}$	3	1.99	3.79	3.02	2.04	2.24
$^3J_{HN,H\alpha}$	2	5.35	4.78	5.50	4.92	5.59
$^3J_{HN,H\alpha}$	3	5.73	5.29	5.92	5.36	5.74
$^3J_{HN,C\alpha}$	2	0.63	0.63	0.63	0.63	0.67
$^3J_{HN,C\alpha}$	3	0.62	0.62	0.62	0.62	0.68
χ^2[l]		0.53±0.02	2.91±0.06	1.08±0.03	0.67±0.02	

[l] Uncertainties on χ^2 values represent one standard error of the mean calculated using a block averaging method.[64] Uncertainties on individual J-coupling constants are omitted for clarity.

Note that the calculated J-couplings depend on the ratios of various backbone conformations, between which transitions may be relatively rare. In our studies of Ala_5, we found that 1.5 μs of aggregate simulation time, which we had previously used in our validation of ff14ipq,[91] did not yield sufficiently precise calculations of the J-couplings. Although the J-couplings may appear to be converged based on their relatively small statistical variances (evaluated using block averaging), these variances may be misleading. For example, we observed what appeared to be small, but statistically significant differences in the J-couplings between simulations run with and without hydrogen mass repartitioning after 1.5 μs of simulation, but these differences ultimately disappeared after 6 μs. The extensive sampling needed to obtain converged J-couplings illustrates the challenge of mapping the conformations of just a few residues using brute-force MD simulation.

3.4.4 α-helices: K19 and $(AAQAA)_3$ peptides

To assess the propensity of ff15ipq to form α-helices, we studied the temperature-dependent behavior of two model α-helical peptides: K19 and $(AAQAA)_3$.[120,121] Both peptides are variants of the motif $(Ala\text{-}Ala\text{-}Xaa\text{-}Ala\text{-}Ala)_n$, in which Xaa is Lys in K19 and Gln in $(AAQAA)_3$; their sequences are listed in Table 3.1. For each peptide, we carried out six 4-μs simulations at different temperatures and monitored the formation of various types of secondary structure. As shown in Figure 3.4, both peptides undergo multiple folding and unfolding events, though our simulations are not sufficiently long to obtain converged estimates of the probability of adopting α-helical conformations. Qualitatively, K19 adopts α-helical conformations for a greater proportion of the simulation than $(AAQAA)_3$, which is consistent with the experimental observation that K19 and $(AAQAA)_3$ are ~40% and ~20% α-helical, respectively, at 300 K.[120,121]

Both peptides transiently form β-sheet contacts, which do not appear to be stable for more than 100 ns, indicating that ff15ipq correctly identifies the favored secondary structures of these peptides.

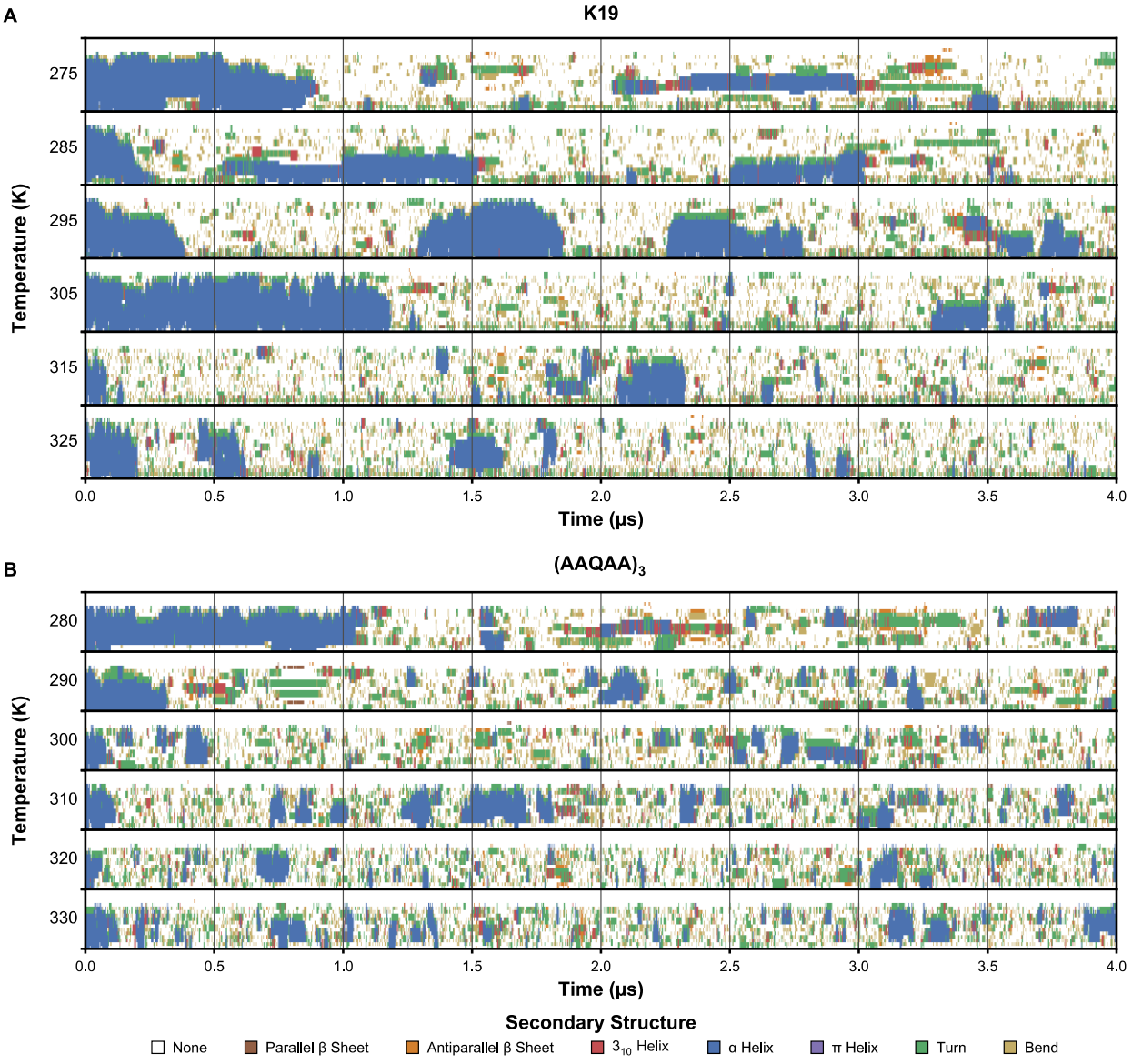

Figure 3.4. Secondary structure of model α-helical peptides K19 (A) and $(AAQAA)_3$ (B) at various simulated temperatures over the course of 4-μs simulations.

While the two peptides differ in length, the observed difference in α-helical stability is likely due to parameters of Gln and Lys residues at the central positions of the peptides. In our umbrella sampling simulations of tetrapeptides, which are too small to form an α-helix (Figure 3.16), we observed a broader, deeper α well for Lys than for Gln, suggesting that the observed difference in α-helical stability between the two peptides has already been "built-in" to ff15ipq at the residue level. In addition, the two residues have different backbone charges: Gln shares its N, H, C, and O charges with the other neutral residues, and Lys shares its charges with the positively charged residues. While the backbone H charges for neutral and positively-charged residues are similar, the backbone O of the positively charged residues is ~0.05 e more negative than that of the neutral residues, which may result in more stable hydrogen bonding.

3.4.5 β-sheets: GB1 hairpin, chignolin, and Cln025 peptides

In order to assess the stability of β-sheet structures in ff15ipq, we simulated three model β-hairpin systems: the GB1 hairpin, the designed peptide chignolin, and its hyper-stable variant Cln025.[122–124] We simulated the GB1 hairpin at six temperatures ranging from 275 K to 325 K, Cln025 at ten temperatures ranging from 280 K to 370 K, and chignolin only at 298 K. Figure 3.5 shows the secondary structures observed during 4-μs simulations of these systems. As with our simulations of the α-helical peptides, our β-hairpin simulations are not sufficiently long to precisely quantify secondary structure stability, though qualitative trends may be identified. As shown in Figure 3.5A, the GB1 hairpin is metastable over the tested temperature range of 275-325 K, and in two of our simulations unfolds and refolds. Our simulations at ≥285 K are in qualitative agreement with experiment, which have indicated that the GB1 hairpin is ~85% folded at 275 K, ~50% folded at 295 K, and ~20% folded at 325 K.[152] However, an anomaly is observed in our 275 K simulation,

in which the GB1 hairpin unfolds after ~200 ns and does not refold. This unfolding event may simply be an artifact of our limited sampling that would disappear were the simulations run to convergence. Alternatively, it may reflect limitations of the SPC/E_b water model at temperatures distant from those at which it was parametrized; while the temperature-dependent behavior of SPC/E_b has not been characterized, to our knowledge, three-point water models including the parent SPC/E water model are known to poorly reproduce the temperature dependence of properties such as density.[90,153,154]

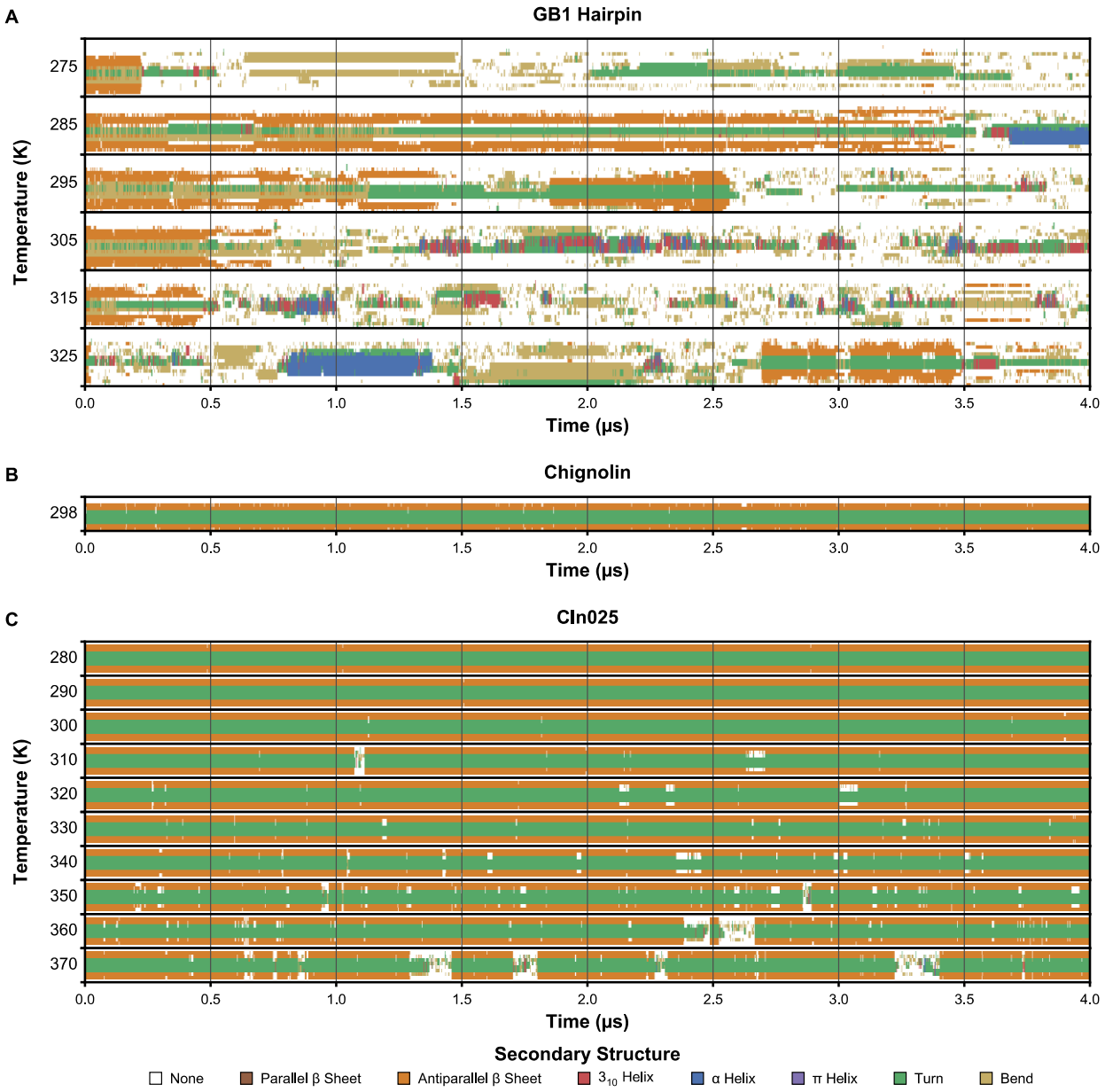

Figure 3.5. Secondary structures of model β-hairpin peptides GB1 hairpin (A), chignolin (B) and Cln025 (C) at various simulated temperatures over the course of 4-μs simulations.

In contrast, our simulations of chignolin and Cln025 suggest that these β-hairpin systems may be more stable than observed experimentally. As shown in Figure 3.5B, chignolin maintains its β-hairpin configuration throughout our 4-μs simulation at 298 K, including two hydrogen bonds in an anti-parallel sheet configuration, while experimentally the peptide is only ~60% folded at

this temperature.[123] Chignolin's hyper-stable variant Cln025 has an experimental melting temperature of 343 K.[124] As shown in Figure 3.5C, in our simulations at temperatures ranging from 280 to 370 K we observe unfolding and refolding events at several temperatures, though the overall folded population is larger than measured experimentally. In particular, Cln025 is >80% folded in our simulation at 370 K, while, experimentally, the peptide is only ~25% folded at this temperature.[124]

As with α-helices, we expect ff15ipq to yield residue-specific propensities in β-sheet stability, though the large difference in sequence between the two tested types of model β-hairpins make comparing them difficult. The aforementioned lack of a clear β well in Asp may destabilize the GB1 hairpin, which contains two adjacent Asp residues, one of which forms part of the anti-parallel β-sheet and the other, the turn. The observed stabilities of chignolin and Cln025 preclude the notion that ff15ipq is biased against β-sheet structure in general. Further studies including additional hairpin sequences and parallel β-sheet structures will be necessary to quantify and mitigate residue-specific biases for future IPolQ force fields.

3.4.6 The Trp-cage miniprotein and globular proteins BPTI, villin, GB3, ubiquitin, binase, and lysozyme

In order to assess the stability of proteins with ff15ipq, we simulated the Trp-cage miniprotein and a series of six globular proteins: BPTI, villin, GB3, ubiquitin, binase, and lysozyme. Extensive experimental data is available for all of these model systems, providing excellent opportunities for validation of our simulations. We carried out 24 μs of aggregate equilibrium simulations of Trp-cage at temperatures ranging from 275 to 325 K and simulations 2-10 μs in duration of the six

globular proteins at temperatures between 298 and 303 K. Details of our simulations are listed in Table 3.1.

The designed miniprotein Trp-cage is central to a long-running computational success story for AMBER force fields. Folding simulations of this miniprotein using the AMBER ff99SB force field have successfully recovered the folded structure, yielded multiple folding and unfolding events, and provided a melting temperature (T_m) of 318 K, which is in reasonable agreement with the experimental T_m of ~315 K.[26,125,155] As shown in Figure 3.6, Trp-cage remained stable in our simulations between 275 K and 295 K, with an average backbone RMSD from the experimental NMR structure of < 1 Å, and unfolded between 305 and 325 K. While our simulations are not extensive enough to obtain precise estimates of the T_m, these results suggest that the T_m is somewhere between 295 K and 305 K, which is slightly lower than experiment. Each unfolding event is marked by an initial shift of the backbone Φ/Ψ of Pro 12 from the α well to the PPII well, followed by the loss of the N-terminal α-helical component of the polypeptide. Notably, in our simulation that was run at 325 K, the protein refolded for ~500 ns, indicating that the folded state is a stable free energy minimum. Thus, despite the extensive reoptimization of the parameter set, the important success of the AMBER force fields in modeling the stability of Trp-cage has been maintained.

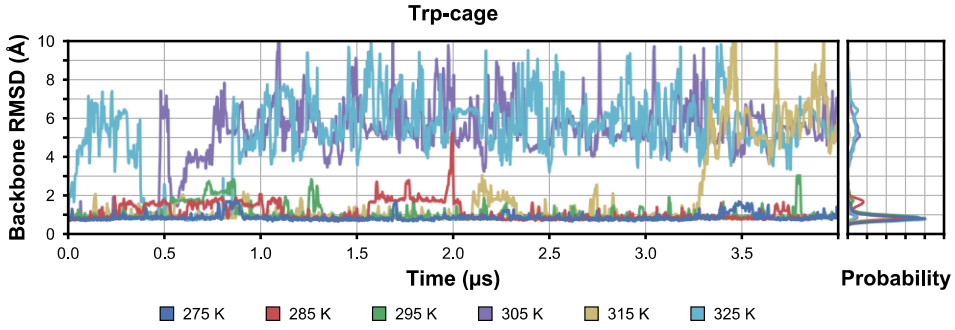

Figure 3.6. Stability of the Trp-cage miniprotein over the temperature range of 275-325 K over the course of 4-μs simulations, as monitored by the backbone RMSD, relative to the experimental NMR structure.

Note that the Pro-rich sequence of Trp-cage allows the opportunity to validate the unique pre-Pro Ψ and Ψ' terms of ff15ipq, because it contains Gly, Arg, and Pro residues that precede Pro. Whenever Trp-cage was folded in our simulations, Gly 11 remained stably in its PPII' well, while Arg 16 sampled broadly across the β and PPII regions without a clear barrier between them. These results are in good agreement with the experimental NMR ensemble, within which the Φ and Ψ backbone torsions of Arg 16 are distributed in a line across these two regions,[125] and with the pre-Pro distributions observed for these residues in the NDRD dataset.[143] Also consistent with both the NMR ensemble and NDRD dataset are the observed distributions of Pro 17 and Pro 18, which strictly maintained their positions in the PPII well even as the protein unfolded.

As shown in Figure 3.7, the overall structures of all six globular proteins – BPTI, villin, GB3, ubiquitin, binase, and lysozyme – remained stable over their entire simulations. The BPTI protein stayed closest to its crystal structure, yielding an average backbone RMSD of 0.7 Å, which may be a result of its three disulphide bonds among a total of 58 residues. All α-helical and β-sheet regions of this protein were retained for the entire simulation. The only notable deviation from the crystal structure of BPTI was observed for Ala 16 and Arg 17, which form the end of the loop preceding the first β-sheet. These two residues, which occupy the α and ξ wells, respectively, in the crystal structure, made temporary excursions of several hundred nanoseconds to alternative conformations before returning to the crystal conformation (Figure 3.17 and Figure 3.18). The villin headpiece subdomain also remained close to its crystal structure, yielding an average backbone RMSD of 1.1 Å. In order to test our parameters for the noncanonical amino acid norleucine (Nle), which was included alongside the canonical residues during development of ff15ipq, we have simulated the fast-folding double-Nle mutant of villin.[131] The Nle residues, both of which are located in the third α-helix, strictly sampled the α-helical well, suggesting that our

parameters for Nle are appropriate. The only significant deviation from the crystal structure of this mutant of villin is a ~0.5 μs excursion made by residues 10-12, which form the end of helix 1 and the loop linking helices 1 and 2, to an alternative conformation different from that observed in the crystal structure (Figure 3.19 and Figure 3.20).

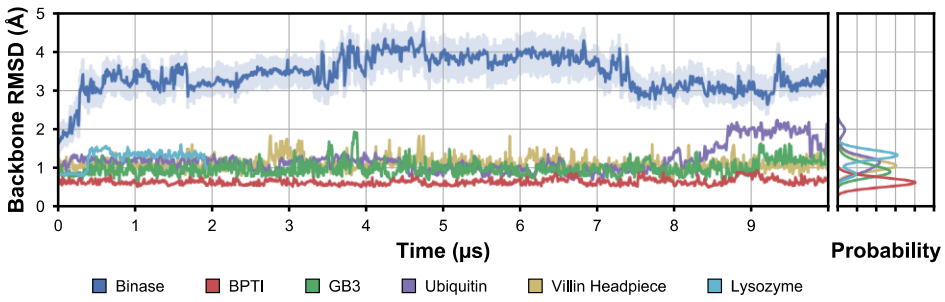

Figure 3.7. Stability of folded proteins over the course of 10-μs simulations as monitored by the backbone RMSD relative to the experimental structures. For binase, the mean RMSD relative to the ensemble of 20 NMR structures is shown along with the range between the minimum and maximum values (light blue shaded region).

Our simulation of GB3 yielded an average backbone RMSD of 1.0 Å from the NMR structure. Three of the residues exhibit significant deviations from the crystal structure: Leu 12, Asp 40, and Thr 55 (Figure 3.21 and Figure 3.22). The conformation of Leu 12, which is located in the turn linking the first and second β-strand, falls precisely between the β and PPII wells in the NMR structure, while both wells were nearly equally sampled in our simulation. As a result of this increased conformational flexibility at Leu 12, adjacent residues also occasionally sampled conformations outside their NMR structure. The presence of a free energy barrier between the β and PPII wells is a necessity for maintaining stable conformations within these wells; it is likely that the forces contributing to the stabilization of Leu 12's unusual conformation in the NMR structure are simply not captured by the functional form of ff15ipq. Consistent with this hypothesis, nearly identical deviations were observed for the ff14SB force field which shares this

functional form.[81] The conformation of Asp 40, which is located in the loop between the α-helix and third β-strand, occupies the ξ well in the crystal structure while other conformations, predominantly β, were sampled in our simulations. Indeed, based on this result and others presented below, the negatively charged residues of ff15ipq completely lack an ξ well. As with Leu 12, the increased conformational flexibility of Asp 40 led to broader sampling by adjacent residues. It is worth noting that GB3 contains two negatively charged residues, Glu 15 and Asp 22, which remained stable in β-sheets, indicating that the limited sampling of this region observed in our umbrella sampling simulations may be overcome within the context of a folded protein. Finally, a notable deviation from the NMR structure occurs for Thr 55 during the last microsecond of our simulation when the antiparallel β-sheet hydrogen bonds between this residue and Val 42 are broken, though the remainder of the β-sheet remains in place. While this deviation may be a transient event, it could also be a consequence of the conformational deviations of the nearby Asp 40.

Similar to GB3, ubiquitin in our simulations exhibited a low overall average backbone RMSD of 1.2 Å from the crystal structure, but significant deviations in certain regions. In particular, transient deviations from the crystal structure were observed for residues 8-11, which form the turn connecting the first and second β-strands (Figure 3.23 and Figure 3.24). While these residues sampled conformations that were different from the crystal structure, their turn conformation was retained throughout 80% of the simulation, and experimental NMR relaxation data suggests that is region is truly flexible.[156] A more significant deviation was observed for Asp 52 and Gly 53, which are located in a loop region; in the crystal structure, these residues both occupy the α well while in our simulation, Asp 52 and Gly 53 shift to the PPII and γ wells, respectively. Unlike the deviations observed in GB3, this shift does not lead to broader sampling

by adjacent residues or appear to otherwise destabilize the protein, suggesting that the observed alternative conformation may simply be erroneously modeled by ff15ipq to be lower in energy than the conformation found in the crystal structure. Finally, after ~8.5 μs of simulation, Glu 34, which is the last helical residue in the central α-helix, shifts to the PPII and β regions, leading to shifts in residues 33-36. Combined with the observations made for Asp 52, this shift suggests that for negatively charged residues, ff15ipq may overstabilize PPII conformations relative to α.

Among the simulated globular proteins, the greatest deviations from the initial structure were observed for binase, with an average backbone RMSD of 3.4 Å from the NMR ensemble of 20 models.[126] The same average RMSD was also obtained with respect to the crystal structure of wild-type binase, which differs in the amino acid sequence at six positions (PDB code: 1GOU).[157] These larger differences are primarily caused by variability in loop regions as noted for the experimental structures,[126,157] and the core structure of binase remained relatively close to the experimental structures with an average backbone RMSD of 1.9 Å. The first loop, which comprises residues 34-39, adopted multiple conformations in our simulation, which is consistent with the NMR ensemble (Figure 3.25 and Figure 3.26). Moreover, this loop adopts the conformation observed in the crystal structure for 75% of the simulation. The second loop comprises residues 56-62, which also sampled broadly in our simulations, consistent with diverse conformations in the NMR ensemble and poorly defined electron density in the crystal structure.[157] Notably, in both our simulation and the NMR ensemble, flexibility in this region extends to Gly 67. This difference may relate to the difference in sequence between the NMR and crystal structure proteins, in which Ser 66 is replaced by Gly and Gly 67 by Ser. The third loop, comprised of residues 76-83, is also broadly sampled in our simulation, and is the source of greatest difference relative to the experimental structures. The final loop is comprised of residues 99-104 and in our simulation we

observe several residues sampling two different conformations, consistent with the observation of two states in crystal structures determined under different conditions.[157]

The largest protein system, lysosome, was simulated for 2 μs, over which it exhibited an average backbone RMSD of 1.2 Å with respect to the crystal structure. As shown in Figure 3.27 and Figure 3.28, the largest deviations were found in the loop comprised of residues 100-104. Residues 101 through 104 adopted an alternative conformation, with the flanking residue Val 99 no longer part of the preceding α helix and Gly 104 part of the following helix. Similar to our observations for ubiquitin, this difference appears to be related to the shift of a negatively charged residue, Asp 101, from the α well to the PPII well.

Four of the the six globular proteins – BPTI, GB3, ubiquitin, and lysozyme – have also been used for validation of previously developed force fields, including ff14ipq, ff14SB, and CHARMM36.[81,87,91] Similar to these force fields, ff15ipq yielded low average backbone RMSD values for these proteins, relative to their initial structures (i.e., ≤1.2 Å). A key point to be considered while comparing our results to those of previous force fields is that advancements in GPU computing over the last several years[79] have enabled us to validate ff15ipq using simulations up to 10 μs, which is up to 10x as long as those used for the other force fields. In particular, ff14SB was validated using sets of four 1-μs simulations,[81] while CHARMM36 was validated using 200-ns simulations.[87] Many of the key deviations we observe do not occur for several microseconds; for example, we observe changes in the C-terminus of GB3 after 9 μs, and in the loops of ubiquitin after 8.5 μs. These deviations are informative and will guide development of successors to ff15ipq, illustrating the utility of long-time scale simulations for force field development.

A particularly appealing feature of the SPC/E_b water model with which we have developed ff15ipq is its ability to more accurately reproduce the rotational diffusion of solvated proteins

relative to water models such as TIP3P and TIP4P-Ew.[96] In order to measure how accurately the combination of ff15ipq and SPC/E_b are able to reproduce rotational diffusion, we branched off sets of ten 200-ns simulations in the microcanonical ensemble (NVE) from our 10-μs simulations in the canonical ensemble (NVT) for GB3, ubiquitin, and binase, thereby avoiding perturbation of the dynamics by the use of a thermostat. As shown in Table 3.3, ubiquitin, GB3, and binase, diffused ~14%, ~15%, and ~22% more slowly than measured experimentally by NMR, respectively. Note that the experimental values were corrected for differences in temperature, isotopic labeling, and solvent D_2O content, potentially introducing error into the comparison of simulated and experimental values.[29] Interestingly, while the errors we obtained were consistent with our test simulations with the AMBER ff99SB-ILDN force field and SPC/E_b (Figure 3.13), we found CHARMM22* and SPC/E_b to yield lower errors of 7%, 6%, and 16%, illustrating the coupling of solute and solvent parameters on the motions of proteins through solution. Since the SPC/E_b water model had been empirically optimized for proteins with the AMBER ff99SB force field, this is no fault of ff15ipq (or CHARMM22*, for that matter), but suggests that improved performance might be obtained by optimizing the protein and solvent models in tandem. Finally, it is worth noting that our use of the Langevin thermostat in the NVT simulations results in ~42%-52% longer rotational diffusion times, relative to those of the NVE simulations, demonstrating, for the first time (to our knowledge), that the use of a thermostat can significantly perturb dynamical properties for proteins and not just for small molecules and polymer chains.[158]

Table 3.3. Rotational diffusion of globular proteins simulated with ff15ipq.

System	Experimental τ_c (ns)[m]	Simulated τ_c, One 10-μs NVT simulation (ns)[n]	Simulated τ_c, Ten 200-ns NVE simulations (ns)[o]
GB3	3.03	4.94±0.02	3.47±0.05
Ubiquitin	4.07	6.67±0.04	4.62±0.08
Binase	5.95	11.06±0.07	7.26±0.18

A major advantage of performing simulations with accurate rotational diffusion is that one can directly calculate NMR relaxation parameters ^{15}N $R_1$and R_2, and ^{15}N-^{1}H heteronuclear NOE, which report on the dynamics of individual residues, and compare these values with experiment. Therefore, we calculated relaxation parameters for GB3 and ubiquitin, for which experimental data are available at five and four magnetic field strengths, respectively (Figure 3.8 and Figure 3.9). Overall, our calculated R_2 values are in excellent agreement with experiment, with average mean absolute percent errors (MAPE) of 8% for GB3 and 9% for ubiquitin. Our R_1 values are also in good agreement, with average MAPE of 10% and 12%, with a consistent offset observed across all residues. However, our heteronuclear NOE values are somewhat poorer agreement, with average MAPE of 22% and 30% for the two systems.

[m] Experimental rotational diffusion measured using NMR relaxation[128,156,218] and corrected for differences in temperature and D_2O content between simulation and experiment.[29]
[n] Uncertainties represent one standard error of the mean calculated from 50 consecutive 200-ns blocks from a single 10-μs simulation.
[o] Uncertainties represent one standard error of the mean calculated from 10 independent 200-ns simulations.

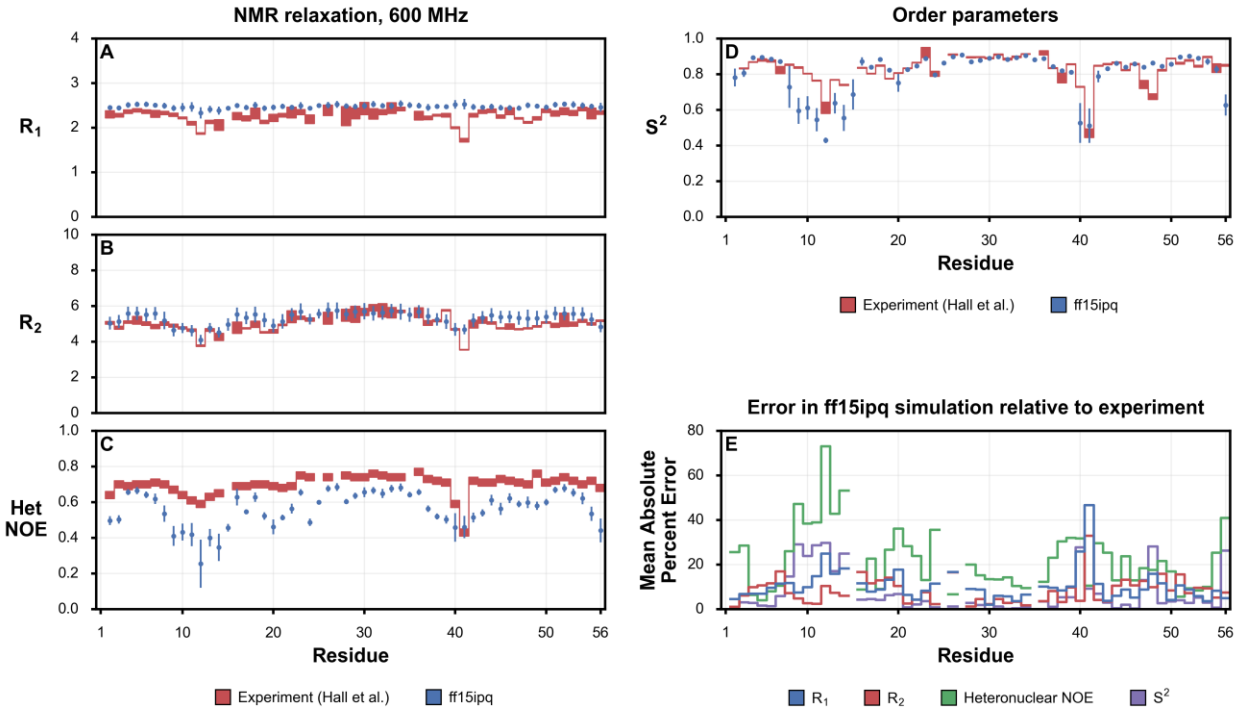

Figure 3.8. Comparison of NMR relaxation and order parameters for the GB3 protein from simulation and experiment. All parameters from simulation were calculated by applying the iRED method[140] to a set of ten 200-μs simulations with the ff15ipq force field, considering the autocorrelation function of each backbone N-H vector to 17 ns, which is five times the overall rotational correlation time (τ_c) of the protein.[159] Experimental R_1, R_2, and heteronuclear NOE data were available at five magnetic field strengths (400, 500, 600, 700, and 800 MHz), and experimental order parameters (S^2) were based on the combined data set from all five magnetic field stengths.[160] At a magnetic field strength of 600 MHz (A, B, C), ff15ipq yields good agreement in R_2, reasonable agreement in R_1 with a consistent offset, and somewhat poorer agreement in heteronuclear NOE; similar agreement is achieved at other magnetic field strengths. Error bars represent 95% confidence intervals; standard errors were obtained by calculating the standard deviation across the ten simulations. Comparison of simulated and experimental S^2 values (D) shows acceptable agreement, with deviations in many of the same regions as those observed for the NMR relaxation parameters. The mean absolute percent error (MAPE) in simulated R_1, R_2, and heteronuclear NOE (E) was calculated by averaging the proportional errors in each parameter across the five magnetic field strengths, while the error in S^2 was calculated relative to the single experimental dataset. Residues with above-average MAPE in R_1 and R_2, including Leu 12, Asp 40, and Gly 41, are discussed in the main text.

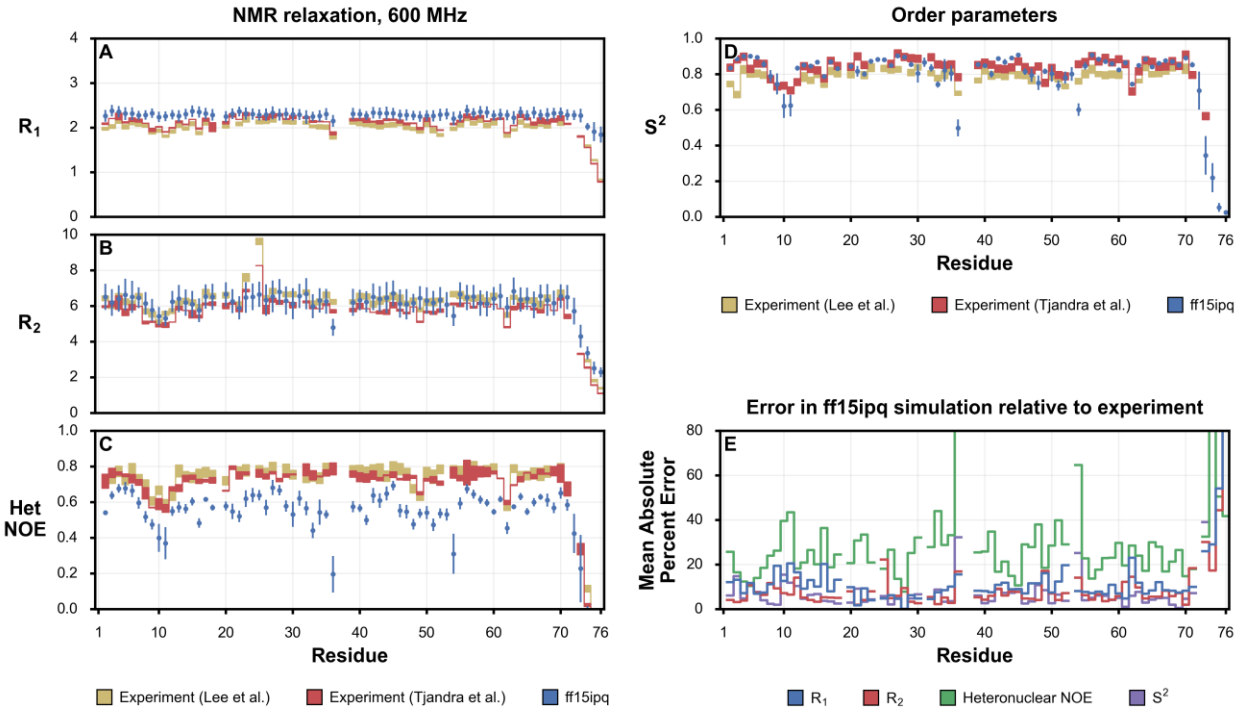

Figure 3.9. Comparison of NMR relaxation and order parameters for the ubiquitin protein from simulation and experiment. All parameters from simulation were calculated by applying the iRED method[140] to a set of ten 200-ns simulations with the ff15ipq force field, considering the autocorrelation function of each backbone N-H vector to 23 ns, which is five times the overall rotational correlation time (τ_c) of the protein.[159] Experimental R_1, R_2, and heteronuclear NOE data were available from Lee *et al.* at four magnetic field strengths (400, 500, 600, and 750 MHz), and from Tjandra *et al.* at 600 MHz, and experimental order parameters (S^2) were available from both groups.[156,161] At a magnetic field strength of 600 MHz (A, B, C), ff15ipq obtains good agreement in R_2, reasonable agreement in R_1 with a consistent offset, and somewhat poorer agreement in heteronuclear NOE; similar agreement is achieved at other magnetic field strengths. Error bars represent 95% confidence intervals; standard errors were obtained by calculating the standard deviation across the ten simulations. Comparison of simulated and experimental order parameters (D) shows acceptable agreement, with deviations in many of the same regions as observed for the NMR relaxation parameters. The mean absolute percent error (MAPE) in simulated R_1, R_2, and heteronuclear NOE (E) was calculated by averaging the proportional errors in each parameter across the four magnetic field strengths, while the error in S^2 was calculated relative to the two experimental datasets. Residues with above-average MAPE in R_1 and R_2, including Lys 11, Asn 25, and Asp 52, are discussed in the main text.

Residues for which our calculated R_1 and R_2 differ by >20% from the experimental values may be due to limitations of our force field. Notably, we find that several such residues are those for which we also observed deviation in sampled backbone conformations relative to the experimental structures. For GB3, Leu 12 and Asp 40 yielded above average errors in R_1 (25%), while Gly 41 had a larger error in both R_1 and R_2 (47% and 33%). Similarly for ubiquitin, Lys 11, and Asp 52 yielded above average errors in R_1 (20%). Also in ubiquitin, Asn 25 yielded an error of 22% in R_2; however, this residue is found experimentally to undergo chemical exchange,[156] fluctuating at time scales beyond those captured by our simulations. Among residues with errors below 20%, one particular trend is apparent: Ile 7, Thr 16, Thr 49, and Thr 51 of GB3 and Thr 12 and Ile 36 of ubiquitin all yielded errors in R_2 of ≥15%, despite having tightly restricted Φ/Ψ sampling consistent with their experimental structures. This trend suggests that ff15ipq may have some discrepancy with the three branched residues that is not apparent when examining only backbone Φ/Ψ preferences, and will be the subject of further study.

3.4.7 Disordered peptides: p53 peptide, S-peptide

In order to test the suitability of ff15ipq for simulating disordered proteins, we focused on two model peptides: the N-terminal, 13-residue peptide fragment of the tumor suppressor p53, and the 22-residue S-peptide fragment of RNase A. Both of these disordered peptides (p53 peptide and S-peptide) only adopt α-helical conformations when bound to their structured partner proteins (MDM2 and S-protein, respectively).[162,163] For each of these peptides, we carried out two 10-μs simulations: one of the isolated peptide and the other of the native peptide-protein complex.

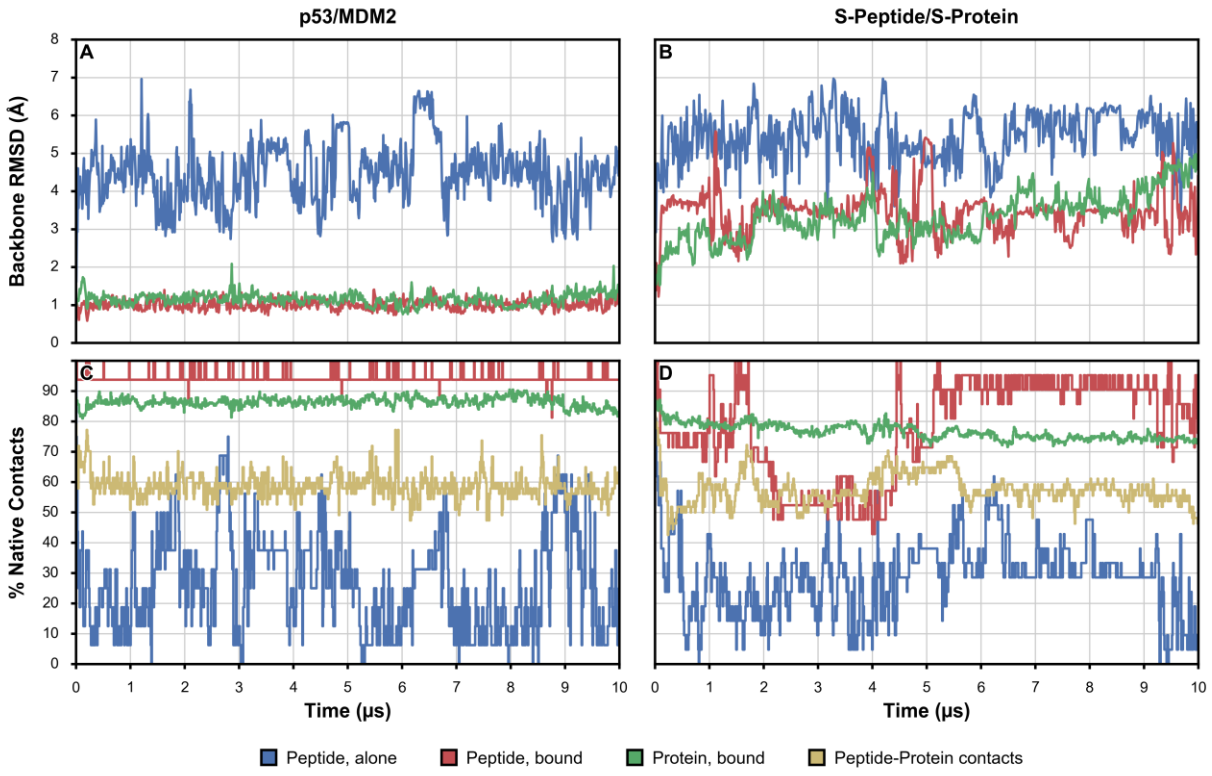

Figure 3.10. Stability of p53 and S-peptide alone and in complex with binding partners MDM2 and S-protein over the course of 10-μs simulations as measured by backbone RMSD relative to the crystal structures (A, B) and the percent of native contacts formed (C, D).

As shown in Figure 3.10, both peptides in their isolated states adopted conformations distant from their partner-bound conformations, sampling a diverse set of conformations with average backbone RMSDs of 5 Å from their corresponding bound conformations in the crystal structures of the peptide-protein complexes and maintaining only ~25% of their native intra-peptide contacts. Furthermore, the p53 peptide only transiently adopted α-helical conformations that resembled its partner-bound conformations (backbone RMSD < 3 Å) with these conformations unfolding within ~200 ns (Figure 3.29). In contrast, the S-peptide did not even transiently sample α-helical conformations resembling its bound state; instead, the peptide formed β-hairpins which persisted for periods as long as 4 μs (Figure 3.32).

In our simulation of the p53/MDM2 complex, the peptide remained stably bound to its partner protein for the entire 10 μs of simulation, with an average backbone RMSD of 1 Å from its bound conformation in the crystal structure (Figure 3.10). Curiously, while most of the intramolecular native contacts of the p53 peptide and MDM2 were retained (~95% and ~85%, respectively), only ~60% of the intermolecular p53/MDM2 native contacts persisted. Examination of the structure showed that many of these contacts lay just below the threshold distance of 5.5 Å between their heavy atoms in the crystal structure. Thus, these contacts were no longer "formed" when slightly different conformations were adopted in our simulations. Throughout our simulation, the p53 peptide retained the α-helical structure of residues 19-25 (Figure 3.29). In contrast, our simulation of the S-peptide/S-protein complex sampled conformations more distant from its crystal structure with average backbone RMSDs of 3 Å for both S-peptide and S-protein. Although the S-peptide partially lost its α-helical structure near the N-terminus from ~1.5 μs to ~5 μs in our simulation (Figure 3.10 and Figure 3.32), the structure reformed and persisted for the remainder of the simulation during which both the S-peptide and S-protein retained most of their intramolecular native contacts (~90% and ~80%, respectively). Similar to the p53/MDM2 complex, the S-peptide/S-protein complex retained ~60% of its peptide-protein contacts throughout our simulation.

Our simulations of these flexible, disordered peptides provide an additional opportunity to validate the backbone conformational preferences of ff15ipq. From our four simulations, we have calculated the backbone Φ/Ψ sampling of the p53 peptide (Figure 3.11) and S-peptide (Figure 3.33). For comparison, we have provided the distributions for the peptide sequences obtained from the NDRD dataset, which accounts for the influence of adjacent residues on each distribution.[143] In our simulation of the p53/MDM2 complex, residues 18-27 of p53 occupied exclusively the

wells observed in the crystal structure, aligning with the observed low RMSD and high percentage of native contacts. Our simulation of the isolated p53 peptide exhibits similarities with the NDRD distribution for most residues, but several inconsistencies are informative: in particular, the neutral residues exhibit reasonable overall agreement and the bulky aromatic residues Phe 19 and Trp 23 closely resemble the NDRD distributions, while Leu 22 and Leu 25 sampled the β-region more extensively than suggested by the NDRD. The negatively charged residues Glu 17 and Glu 28 sampled consistently with the NDRD, while Asp 21 missed sampling in the ξ region. The sole positively charged residue, Lys 24, sampled the β-well more extensively than suggested by the NDRD.

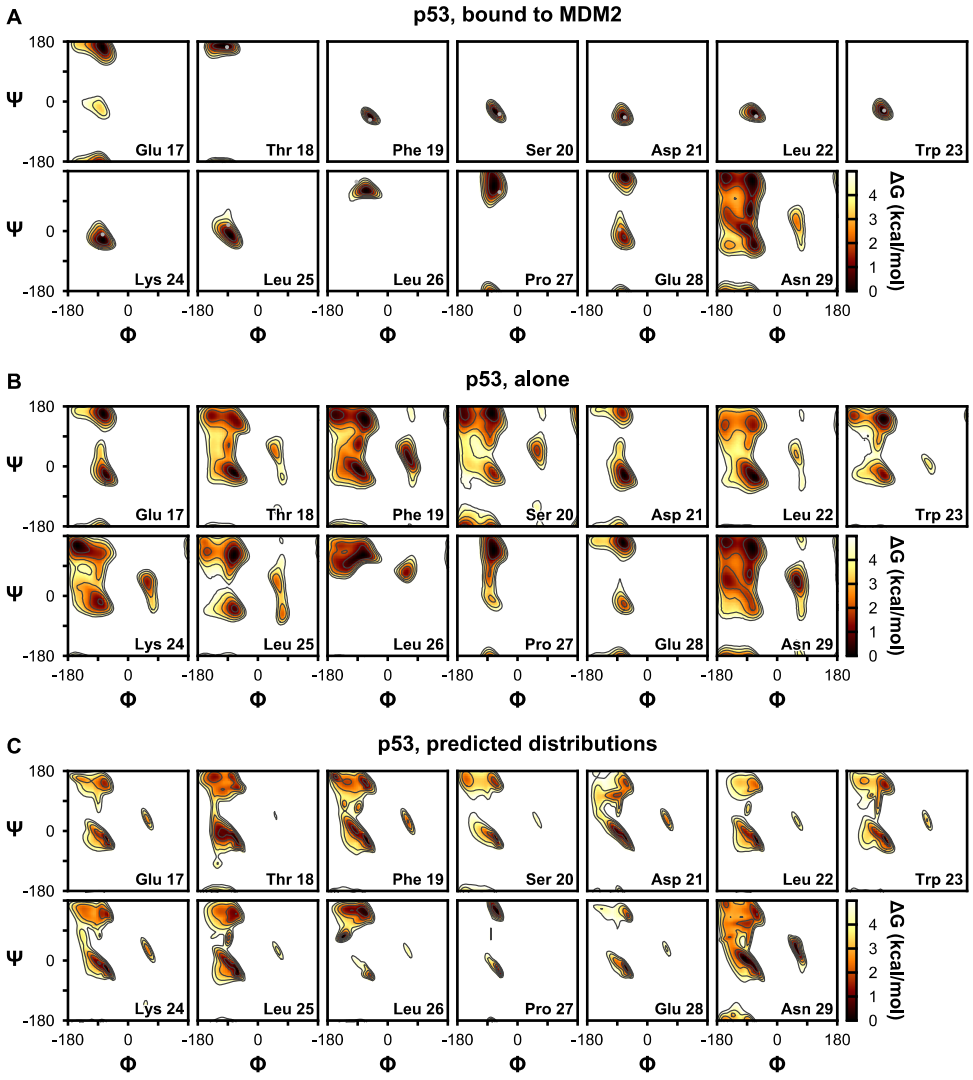

Figure 3.11. Backbone conformational sampling of the disordered p53 peptide observed in 10-μs simulations in complex with the MDM2 protein (A) and alone (B). For comparison are shown distributions for the p53 sequence obtained from the Neighbor-Dependent Ramachandran Distribution (NDRD) dataset,[143] derived from conformations observed in the loops of solved structures (C).

In our simulation of the S-peptide/S-protein complex, the crystal conformations were retained for most of the simulation (Figure 3.33). As described above, several residues near the N-terminus left the α-helical well, and the adjacent Thr 3, which is not helical in the crystal structure, eventually joined the helix as it reforms. The clearest difference from the crystal structure is found

for Asp 14; This residue occupies the ξ well in the crystal structure, which is not present for Asp in ff15ipq, causing it to adopt a PPII conformation. Within the S-protein, residue Gln 60 is notable for its uncommon 'plateau' conformation ($\Phi \approx -100°$, $\Psi \approx -130°$),[133,164] which was retained throughout our simulation (Figure 3.35). In our simulation of the isolated, unbound S-peptide, the formation of long-lived β-hairpin structures prevented us from obtaining converged conformational preferences for comparison with the NDRD distributions, despite the long duration of the simulations (10 μs).

Taken together, the above results indicate that ff15ipq can reliably predict disorder as well as order for peptides that fold upon binding their partner proteins. These encouraging results are worth pointing out since ff15ipq was not specifically parametrized for disordered peptides/proteins, as is the case for contemporary force fields such as ff03w and its subsequent variants.[85,165] As shown in Figure 3.1, both ff03 (whose atomic charges and radii are shared by ff03w) and ff15ipq are able to reliably model propensities of salt bridge formation, which can be critical for such systems that are rich in polar and/or charged residues. Thus, ff15ipq is a reasonable alternative to ff03w for the simulation of disordered peptides/proteins.

3.5 DISCUSSION

Since the establishment of the "one atom, one site" model of all-atom fixed-charge force fields over 20 years ago, several major lineages of protein force fields and countless branches have been developed through cycles of validation and refinement. In this work, we present the ff15ipq force field, the latest in the AMBER IPolQ lineage, and validate its accuracy by >200 μs of aggregate MD simulation. The distinguishing features of ff15ipq are (i) a charge set that accounts directly

for water induced polarization, (ii) the incorporation of two related charge sets for creating new force fields based purely on *ab initio* calculations, (iii) the scope of the parameter optimization, including backbone angles alongside torsions, and (iv) the degree of automation and transferability of the methods to other regions of chemical space. Our simulations suggest that ff15ipq yields reasonable salt bridge propensities, maintains secondary structures and globular protein folds on the μs time scale, predicts order as well as disorder in protein structures, and yields strong agreement with NMR J-couplings and relaxation rates. However, even with this extensive amount of validation, several unconverged results remain and will be explored further using enhanced sampling techniques such as replica exchange[166,167] or recent variants[168,169] of the weighted ensemble path sampling strategy.[170] Here, we will discuss the origins of ff15ipq and its current trajectory.

A major motivation for creating ff15ipq was to address concerns about the overstabilization of salt bridge interactions by ff14ipq, a limitation shared by other contemporary force fields.[56] We addressed these concerns by abandoning the mixed Lennard-Jones radii of ff14ipq and instead increasing the radii of polar hydrogen atoms bonded to nitrogens in both the protein backbone and side-chains to yield more accurate salt bridge propensities. The atomic charges of the force field were then rederived by applying the automated machinery which had created ff14ipq, exchanging the TIP4P-Ew water model for SPC/E_b. Finally, all torsion and selected backbone angle parameters were refit to a QM dataset over four times as large as that used for ff14ipq. In doing so, we found that angle optimization was essential for recovering reliable results with our more extensive dataset.

While the angle optimization feature and other particulars of the torsion fitting are subjects of ongoing development in our force field engine mdgx, our results with ff15ipq indicate that the

workflow illustrated in Figure 3.12 is a viable approach to creating new force fields. Each step entails additional layers of details in order to address practical considerations such as infinite electrostatics or the forms of molecular mechanics basis functions and the shape of the target energy surface. We have addressed each of these issues in Sections 3.2 and 3.3 as well as in prior publications,[33,91] but the synthesis of all these details reflects the physical arguments behind the IPolQ charge derivation.

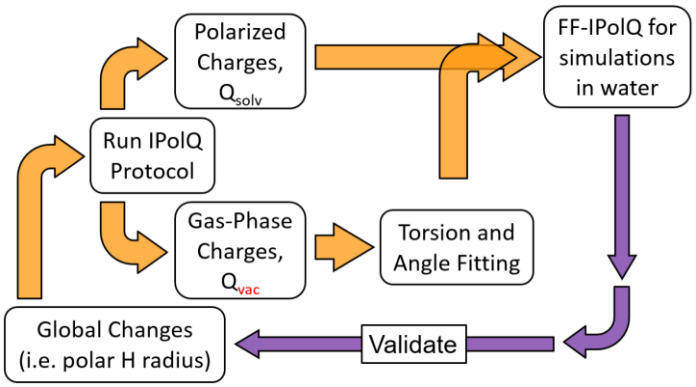

Figure 3.12. IPolQ force field development workflow. Starting from an existing model, selected global changes are optionally first applied to obtain an initial model for optimization. The IPolQ charge deviation protocol is then used to fit a pair of atomic charge sets for the vacuum (Q_{vac}) and solution (Q_{solv}) phases. The vacuum-phase charges are used to fit parameters for bonded terms to vacuum-phase QM targets, and these parameters are subsequently paired with the solution-phase charges to yield a complete force field for solution-phase simulations. The force field is then validated through extensive MD simulation, informing future development.

This approach should be viewed in context with the contemporary Force Balance approach, which also performs sweeping optimization of hundreds of parameters simultaneously.[90] Unlike our MM-minimized conformations, Force Balance typically considers conformations QM-minimized at the same level of theory at which the target QM energies are calculated, though this is not a strict requirement. Going beyond the capabilities of mdgx, Force Balance includes numerous non-linear optimization methods and offers the capability to incorporate results from

sources beyond QM single-point energies, including *in vitro* experiments, directly into the parameter optimization. In the future, such diverse targets might be paired with the IPolQ method, for example by using the vacuum charge set (Q_{vac}) for comparison with vacuum-phase QM data, but to use the polarized charge set (Q_{solv}) in simulations for comparison to experimental results in solution.

It is rather remarkable that a viable protein force field can be produced in months, almost entirely from QM data. Also noteworthy is the fact that features such as angle optimization and generational refinement, which had incremental but definite effects on the accuracy of data fitting, could be so influential in the final result. One way of considering the remaining error in our MM model is to partition it between two sources: bonded and nonbonded interactions. The improvements in ff15ipq that were obtained relative to ff14ipq, whose nonbonded parameters are of similar accuracy, resulted from optimization of angles and the branching of bonded parameters. While the inclusion of anharmonicity in bond and angle stretching or a spline-based treatment of torsion cross terms (CMAP) may reduce errors further,[67,82] greater improvements might be accomplished in the non-bonded interactions.

For this reason, the next planned advance of the AMBER IPolQ force field lineage is to improve the accuracy of electrostatic interactions by making liberal use of virtual charge sites. The "one atom, one site" paradigm used for ff15ipq which was established several decades ago appears sufficiently accurate for most purposes, economical by construction, and thoroughly optimized in existing MD engines. Models with significant numbers of virtual charge sites are presently in the process of becoming established. These models offer improved accuracy for various chemistries with clear physical motivations, accompanied by a more modest increase in computational cost than afforded by polarizable functional forms.[171,172] Future AMBER IPolQ development at the one

site per atom level will continue to explore the applicability of our methodology to the chemical space of other biologically important molecules, including nucleic acids, carbohydrates, and small molecules. Further ahead along the path lie multi-site IPolQ models which will push the mimicry of QM potential energy surfaces – within the confines of a non-polarizable model – to new levels.

3.6 ACKNOWLEDGEMENTS

This work was supported by a University of Pittsburgh Andrew Mellow Fellowship (to K. T. D.), NIH grant 1RO1GM115805-01 (to L.T.C.), NIH grant GM45811 (to D.A.C), and NIH grant RO1GM080642 (to A.M.G.). Computational resources were provided by NSF XSEDE award MCB-100109 for the use of TACC's Stampede; D.E. Shaw Research, NRBSC, PSC, and BTRC for Multiscale Modeling of Biological Systems (MMBioS) through NIH grant P41GM103712-S1 for the use of Anton; NSF MRI award CNS-1229064 for a shared cluster at the University of Pittsburgh's Center for Simulation and Modeling; and private clusters in the Case laboratory.

3.7 SUPPORTING INFORMATION

3.7.1 Supporting methods

3.7.1.1 Simulations of salt bridge formation using the CHARMM Drude-2013 polarizable force field

Systems were initially prepared with the CHARMM36 force field using psfgen, [87,173] equilibrated using NAMD 2.10.0,[109] and uploaded to the CHARMM-GUI server to obtain Drude-2013 topologies.[174,175] The Drude-2013 systems were then subjected to another round of energy minimization and equilibration prior to production simulation. The equilibration process for these systems involved three stages of unrestrained simulation in which a 1-ps simulation was first performed at constant pressure using a 0.1-fs time step, followed by a 20-ps simulation at constant pressure using a 0.5-fs time step, and finally a 1-ns simulation at constant pressure using a 1-fs time step. Production simulations were then carried out for 100 ns at constant pressure using a 1-fs time step. Throughout equilibration and production temperature was maintained at 298 K using a dual Langevin thermostat (frictional constant of 1 ps^{-1}) and 1 K for Drude particles (frictional constant of 20-ps^{-1}),[110] while pressure was maintained at 1 atm using a Langevin piston barostat (piston period and decay time of 200 fs and 100 fs, respectively).[176] Van der Waals interactions were smoothly switched off between 8 and 10 Å, while short-range electrostatic interactions were truncated at 10 Å and long-range electrostatic interactions were calculated using the particle mesh Ewald method.[61] To enable a 1-fs time step, a hard-wall barrier was applied to restrict the maximum distance between Drude particles and their hosts to 0.2 Å,[107] and bonds to hydrogen were constrained to their equilibrium lengths using the M-SHAKE algorithm.[60]

3.7.1.2 Simulations of salt bridge formation using the AMOEBA polarizable force field

Systems were built using TINKER and converted to AMBER format using the AMBERTools 15 software package.[93,177] Each system was subjected to energy minimization followed by a 20-ps simulation at constant temperature and a 1-ns simulation at constant pressure. Production simulations were then carried out for 50 ns at constant pressure. Throughout equilibration and production temperature was maintained at 298 K using a Langevin thermostat (frictional constant of 1 ps^{-1}), while pressure was maintained at 1 atm using a Berendsen barostat (time constant of 1-ps).[178] Van der Waals interactions were truncated at 12 Å, while short-range electrostatic interactions were truncated at 7 Å and long-range electrostatic interactions were calculated using the PME method.[61] Dynamics were integrated using a 1-fs time step.

3.7.1.3 Derivation of amino acid side-chain analogue parameters with AMOEBA

Parameters that are consistent with the AMOEBA polarizable force field were derived for imidazolium, guanidinium, and acetate using the software Poltype.[111] The structures of these side-chain analogues were optimized at the HF/6-31G* level of theory, followed by density calculation at the MP2/6-311G** level of theory using the Gaussian 09 software package.[179] Multipoles were fit using Stone's distributed multipole analysis as implemented in the GDMA program,[180,181] and refined via fitting of the electrostatic potential using a convergence criterion of 0.1 kcal/(mol $electron^2$). Van der Waals radii and well-depths were assigned based on both the element and valence orbitals of each atom, while atomic polarizabilities were assigned based solely on the element. Parameters for bonded interactions were selected based on the analogues' chemical connectivity from a database of small-molecule parameters.[111]

3.7.1.4 Simulations of GB3, ubiquitin, and binase with AMBER ff99SB-ILDN and CHARMM22*

Systems for the simulations reported in Figure 3.13 were prepared and equilibrated using the Desmond 3.0.1.0 software package.[57] Each system was subjected to energy minimization followed by a 20-ps equilibration at constant temperature, and a 1-ns equilibration at constant pressure. Temperature was maintained at 298 K and pressure at 1 atm using the Martyna-Tobias-Klein thermostat and barostat (time constants of 1 ps and 2 ps, respectively).[59] To enable a 2-fs time step, bonds to hydrogen were constrained to their equilibrium values using the M-SHAKE algorithm.[60] A short-range nonbonded cutoff of 10 Å was used, and long-range electrostatics were calculated using the particle mesh Ewald (PME) method.[61]

Production simulations were carried out for 1 μs at constant pressure using a 64-node Anton special-purpose supercomputer and the Multigrator integrator.[25,182] Temperature was maintained at 298 K using the Nosé-Hoover thermostat and pressure at 1 atm using the Martyna-Tobias-Klein barostat (time constants of 1 ps). To enable a 2.5-fs time step, bonds to hydrogen were constrained to their equilibrium values using the M-SHAKE algorithm.[60] Van der Waals and short-range electrostatic interactions were truncated at 10 Å; long-range electrostatic interactions were calculated using the Gaussian split Ewald method,[63] and were updated every third time step.

3.7.2 Supporting figures

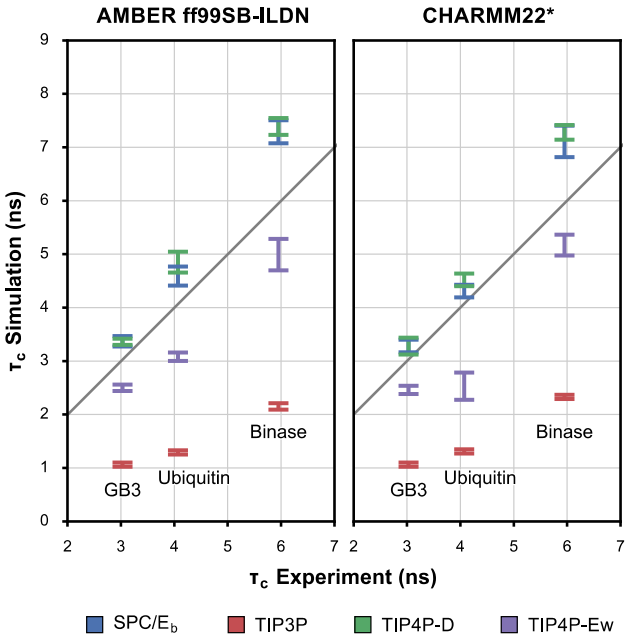

Figure 3.13. Comparison of simulated vs. experimental rotational correlation times τ_c of GB3, ubiquitin, and binase. Simulated τ_c values were obtained using two different force fields (AMBER ff99SB-ILDN and CHARMM22*)[28,32] and four different water models (SPC/E_b, TIP3P, TIP4P-D, and TIP4P-Ew).[27,70,96,102] Experimental τ_c values were measured using NMR relaxation and corrected for differences in temperature and D_2O content between simulation and experiment.[29,156,160,183] Error bars represent 95% confidence intervals; standard errors were obtained by dividing each 1-μs simulation into five 200-ns blocks and calculating the standard deviation.

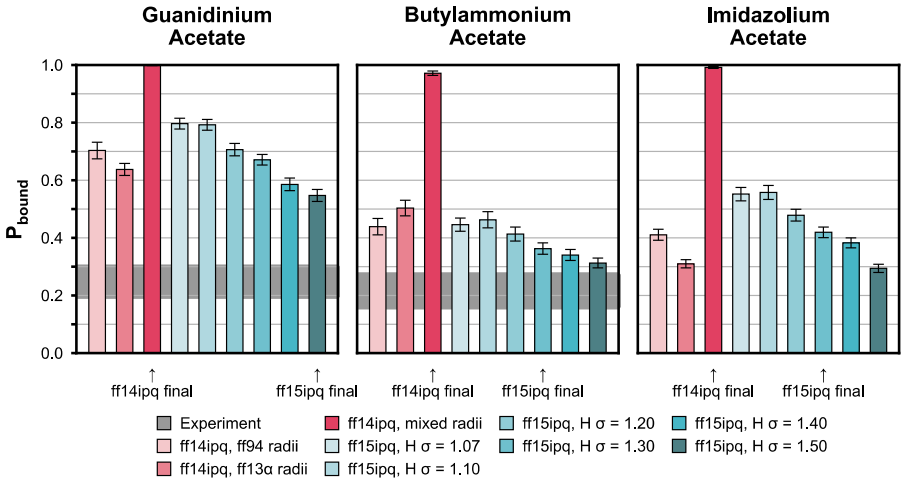

Figure 3.14. Probability of binding (P_{bound}) between acetate and one or more molecules of three cationic side-chain analogues using the ff14ipq and ff15ipq charge sets paired with different Lennard Jones radii. The ff13α radii tested with ff14ipq included an increase in the σ of the carboxylate oxygen (type 'O') and a decrease in the ammonium nitrogen (type 'N') relative to ff94;[33] the final version used 'mixed radii' in which the ff13α σ were applied only to interactions with water.[91] The ff14ipq force field was simulated with the TIP4P-Ew water model,[27] while ff15ipq was simulated with SPC/E_b.[96] The P_{bound} values corresponding to the experimentally-determined K_A values of guanidinium acetate and butylammonium acetate are depicted as horizontal gray bars;[51,69] no experimental value is available for the imidazolium acetate system. Error bars represent 95% confidence intervals calculated using a block averaging method.[64]

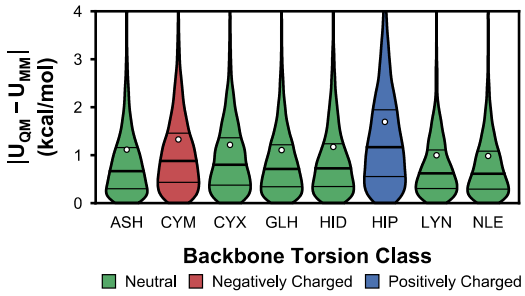

Figure 3.15. Distributions of errors in molecular mechanical energies U_{MM} relative to their quantum mechanical targets U_{QM} for amino acid dipeptides representing alternative protonation states (Ash, Cym, Glh, Hid, Hip, Lyn), the disulphide form of cysteine (Cyx), and the noncanonical amino acid norleucine (Nle). 25th, 50th, and 75th percentiles are represented by horizontal lines, and root mean square errors are represented by white circles. Each dataset is colored by its corresponding backbone torsion class.

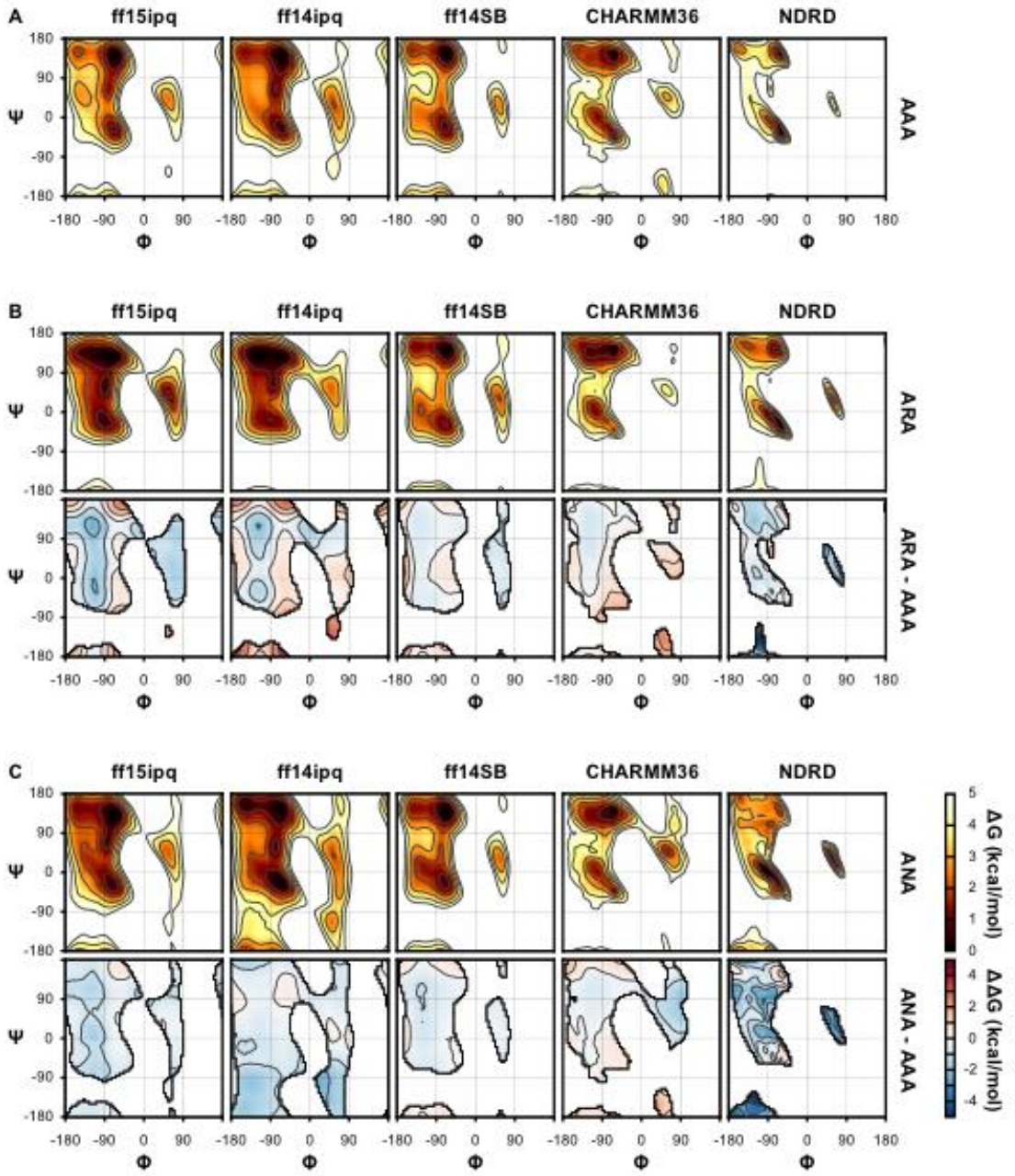

Figure 3.16. Comparison of residue-specific conformational preferences of the central residue of Ace-Ala-Xaa-Ala-Nme tetrapeptides observed in umbrella sampling simulations with distributions of Ala-Xaa-Ala obtained from the Neighbor-Dependent Ramachandran Distribution (NDRD) dataset.[143] Simulations were performed using four different force fields (AMBER ff15ipq, ff14ipq, ff14SB, and CHARMM36),[81,87,91] which were each paired with either the water model with which they were derived or that with which they are most-commonly used. For each system, both the absolute free energy as well as the difference in free energy relative to Ala-Ala-Ala is shown. Regions in which the free energies of both Ala-Xaa-Ala and Ala-Ala-Ala are greater than 5 kcal/mol are omitted for clarity.

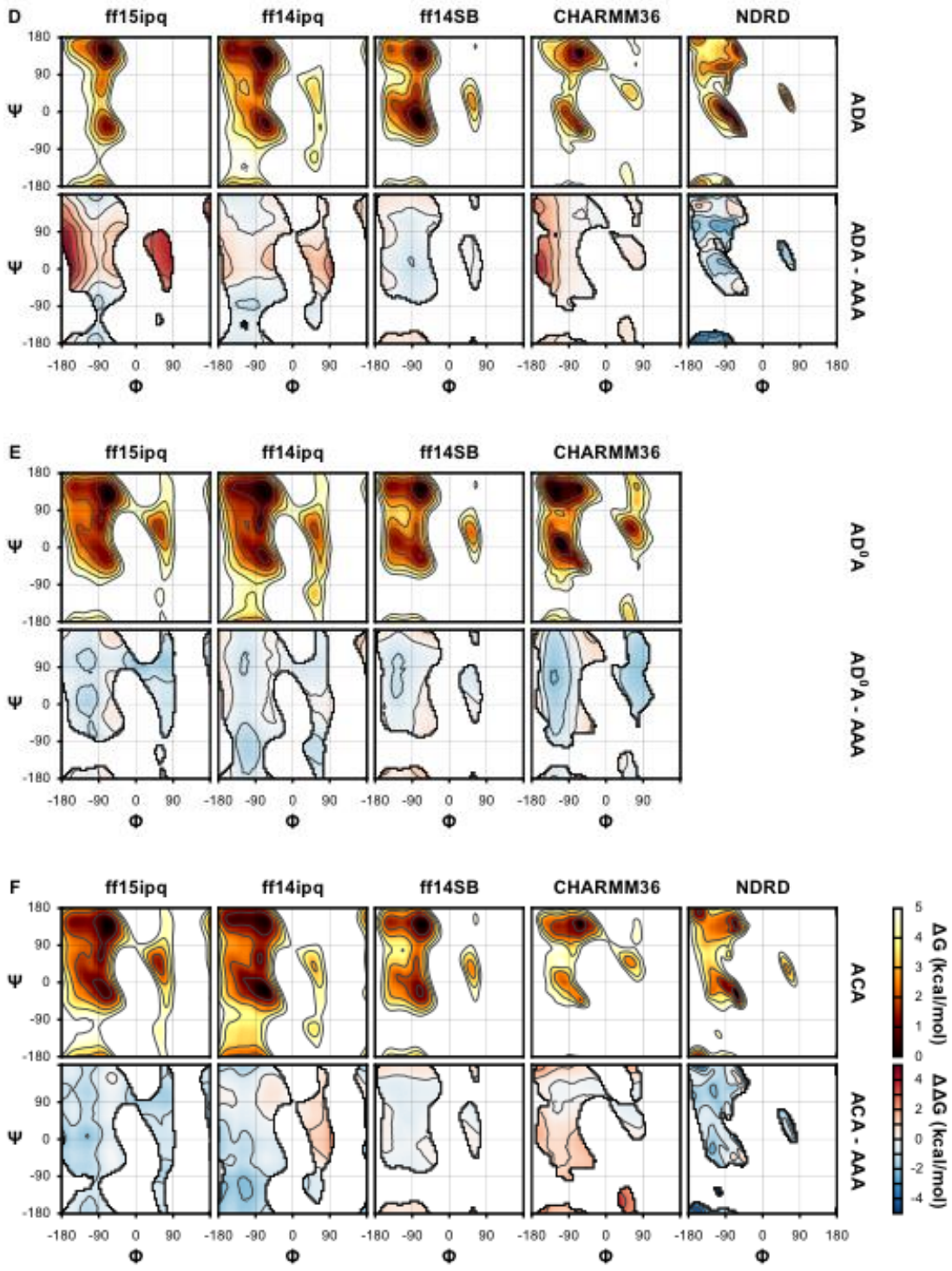

Figure 3.16 (continued).

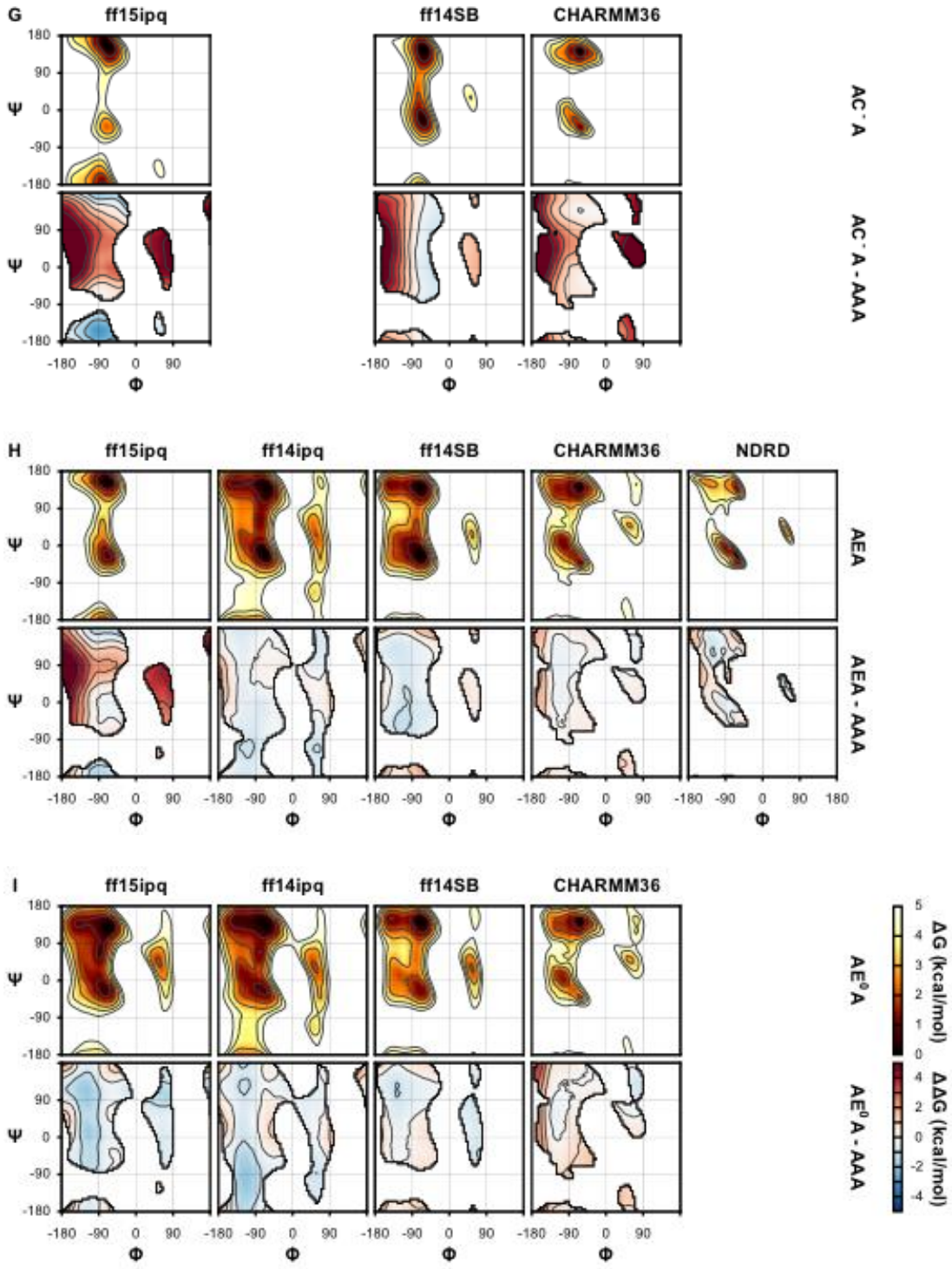

Figure 3.16 (continued).

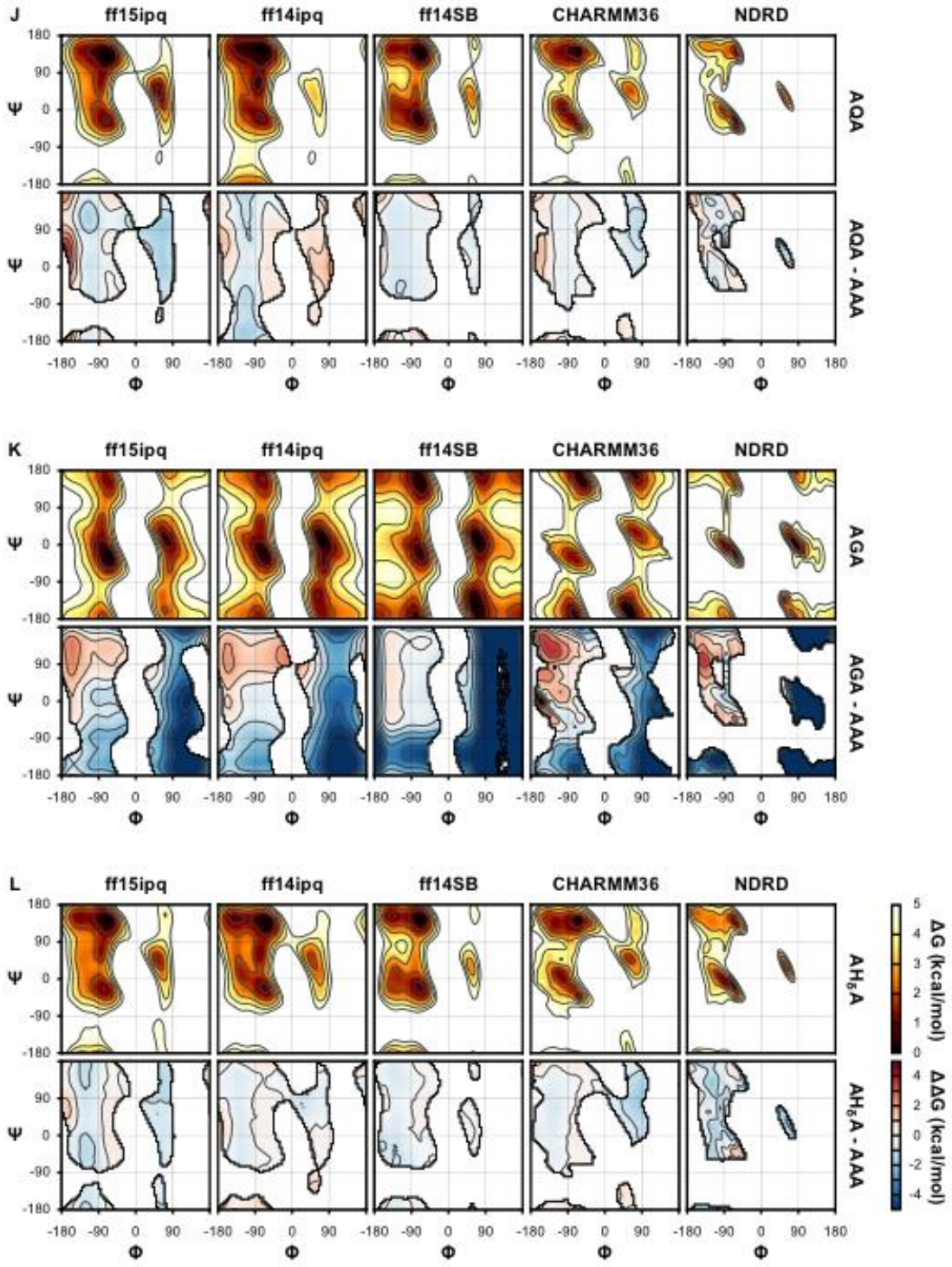

Figure 3.16 (continued).

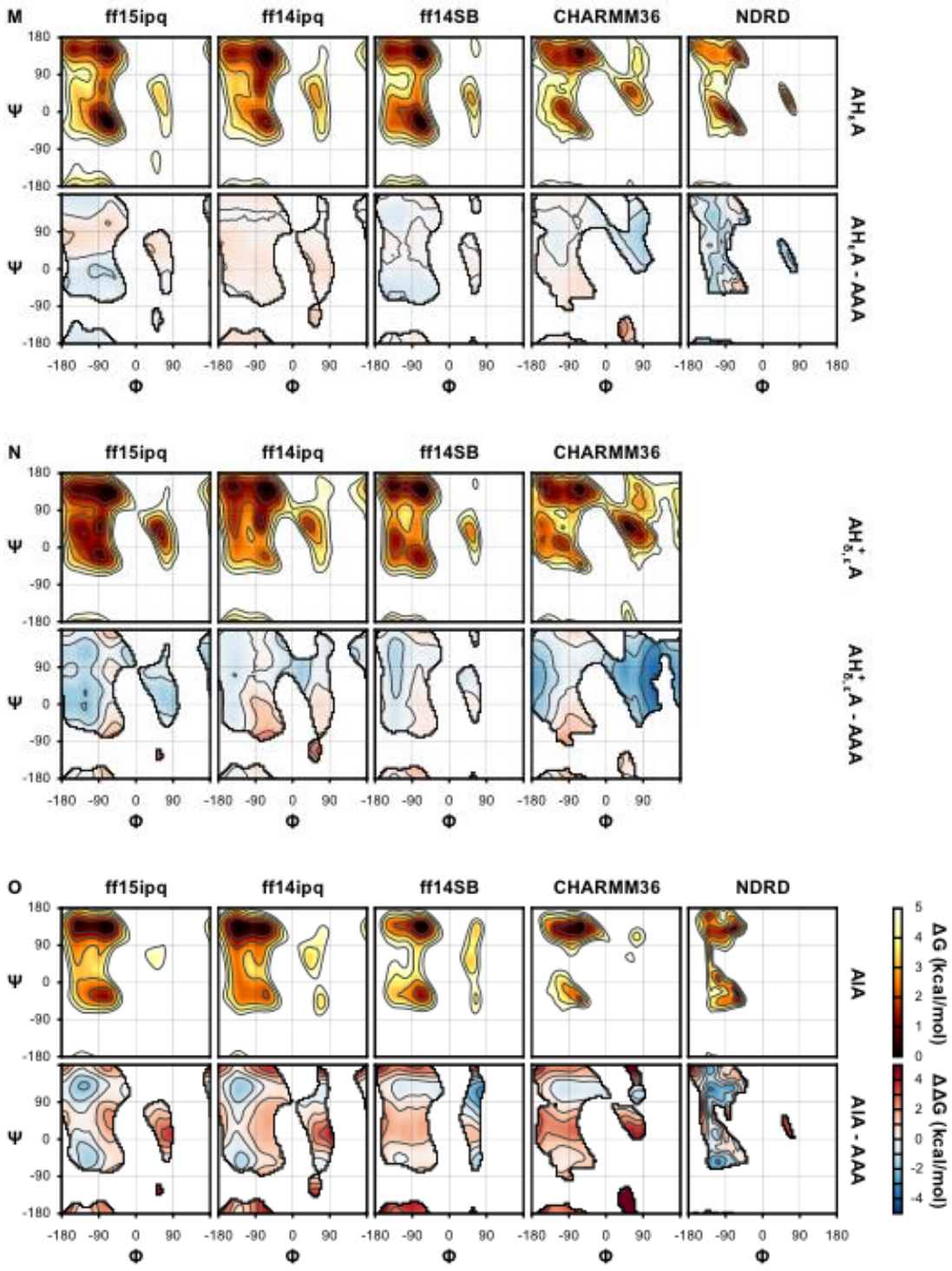

Figure 3.16 (continued).

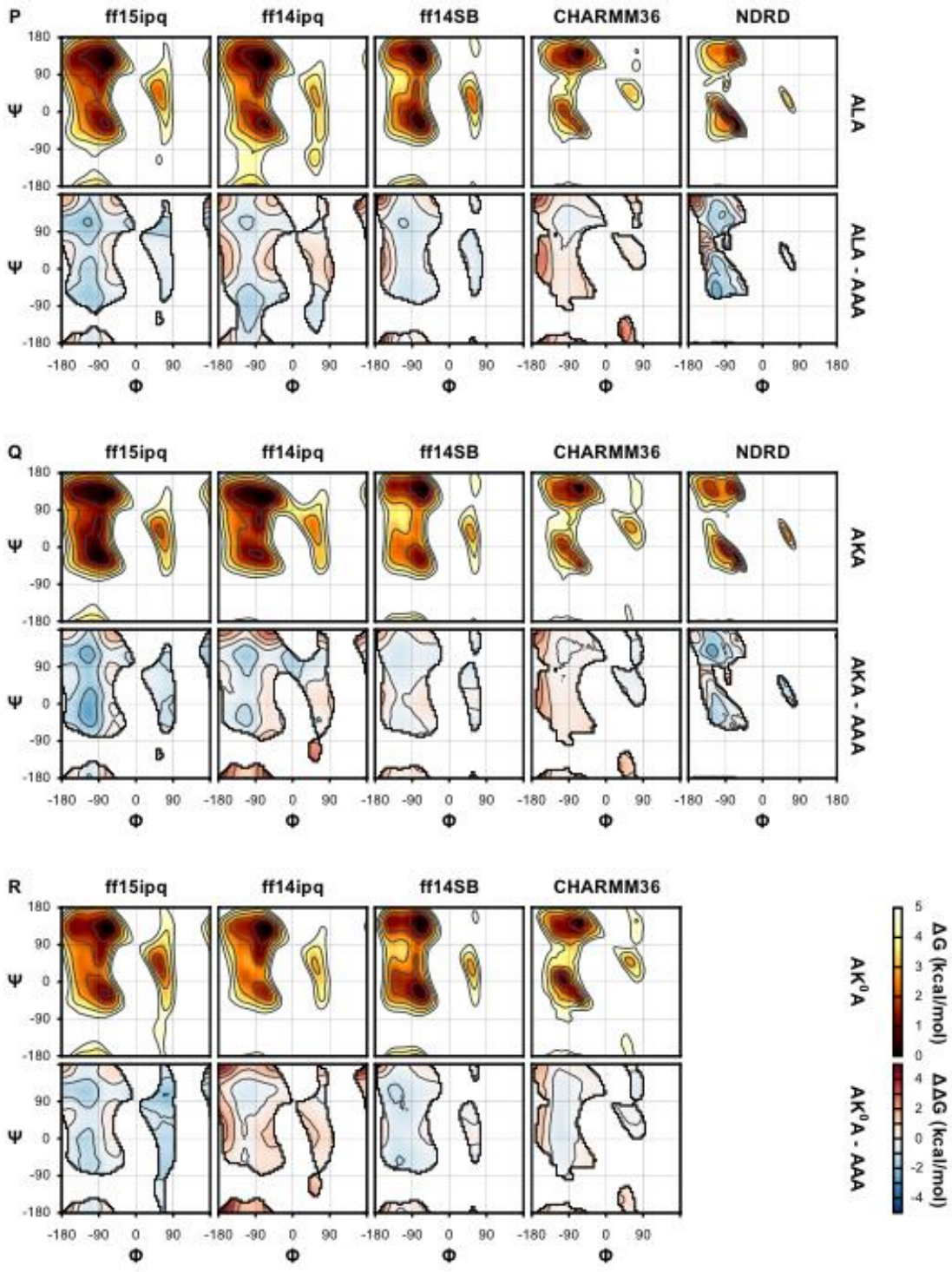

Figure 3.16 (continued).

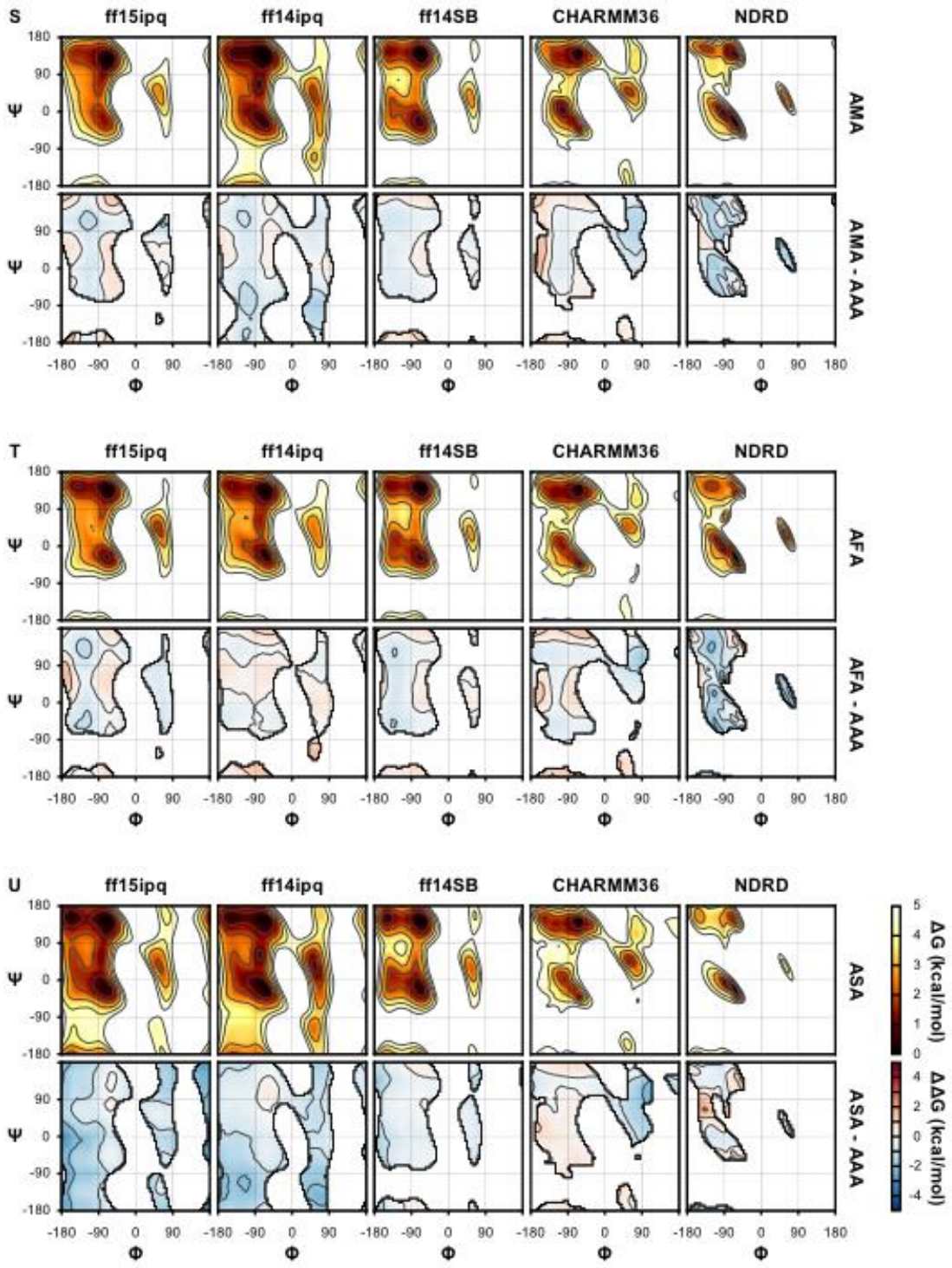

Figure 3.16 (continued).

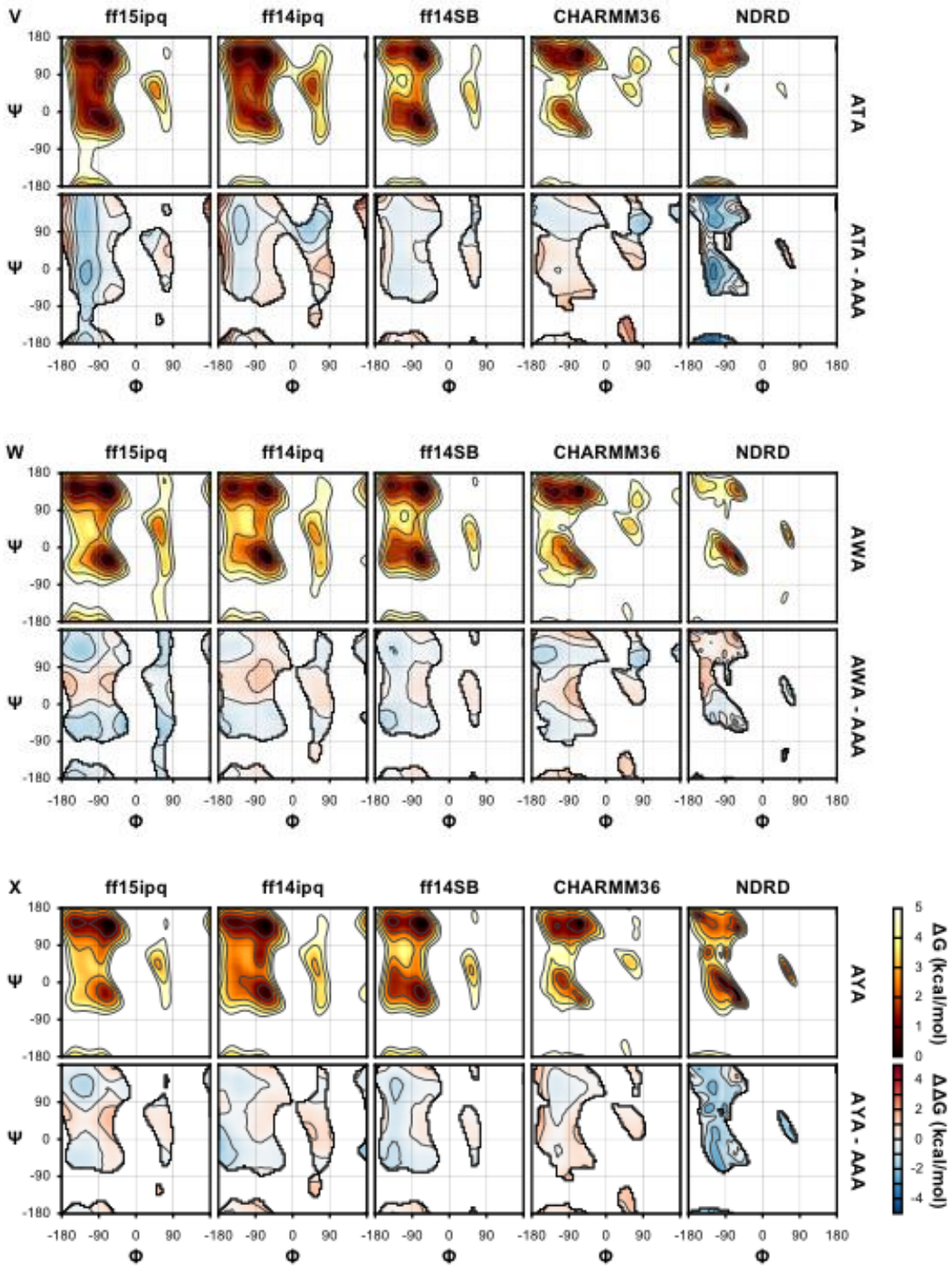

Figure 3.16 (continued).

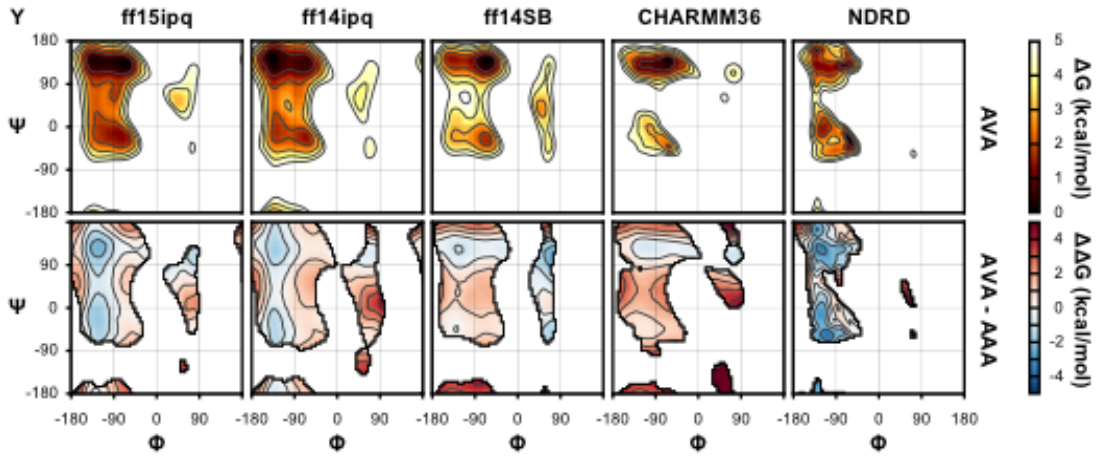

Figure 3.16 (continued).

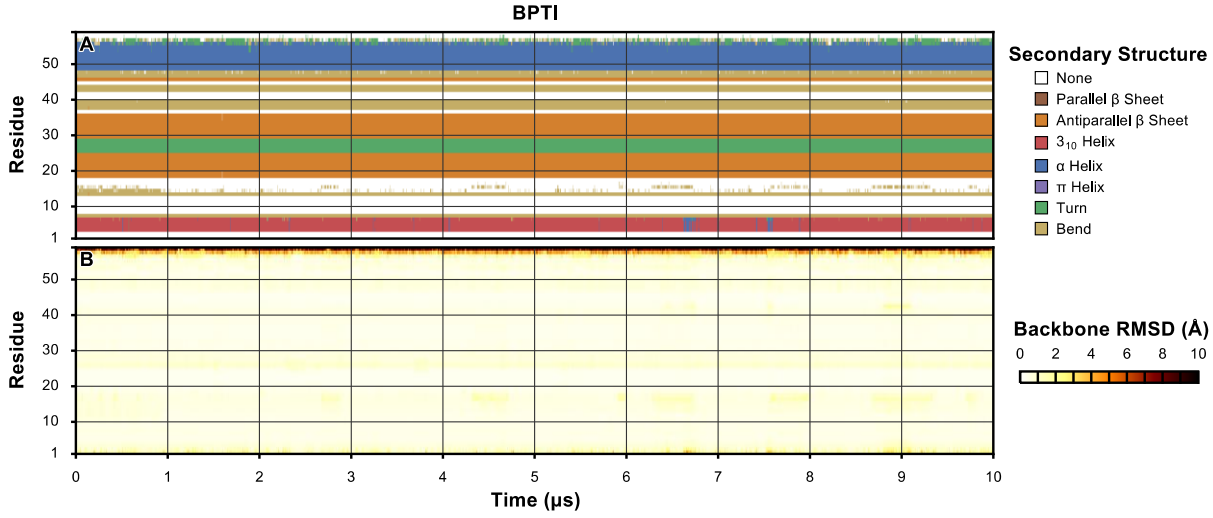

Figure 3.17. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 5PTI)[127] (B) of BPTI observed over the course of a 10-μs simulation.

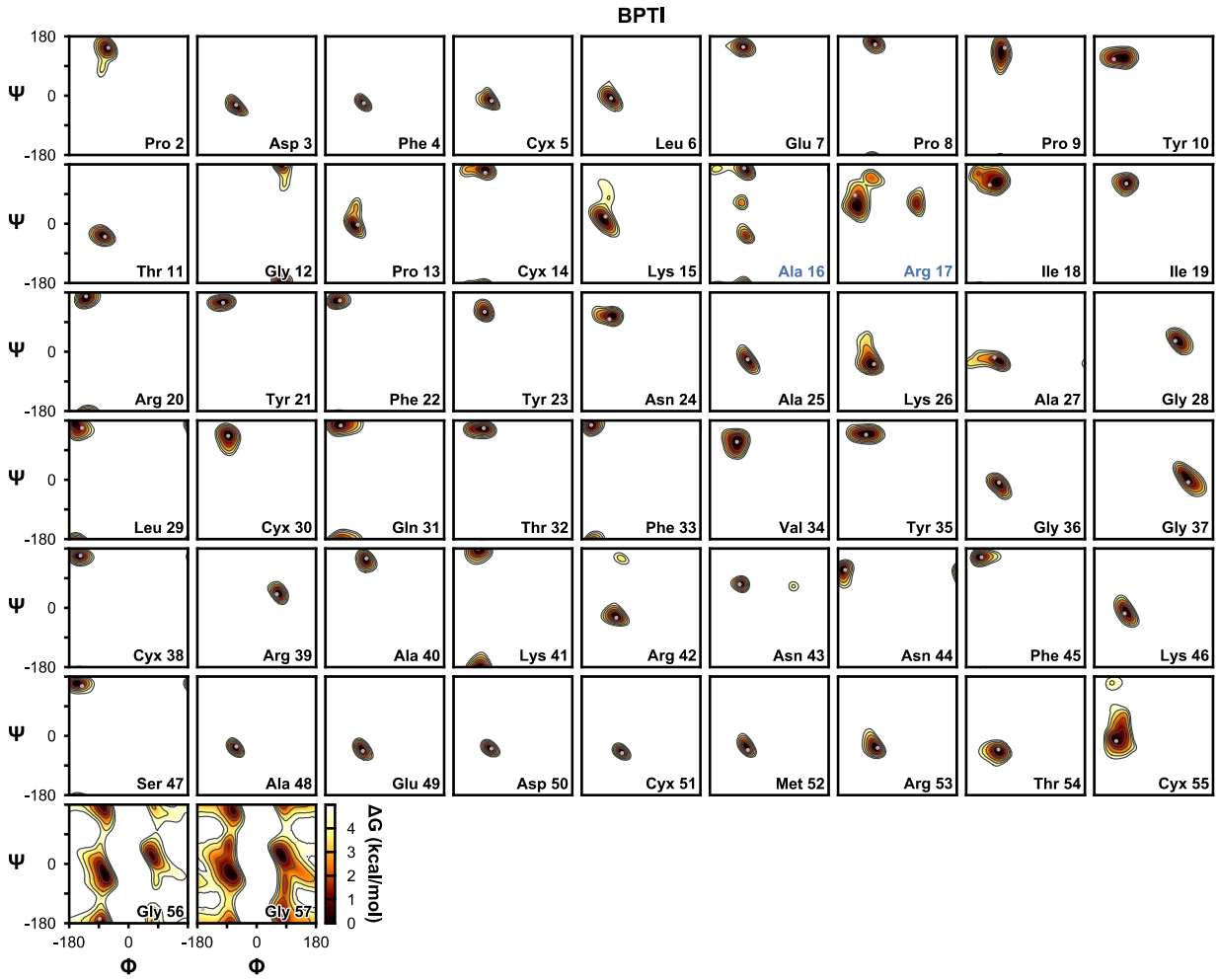

Figure 3.18. Backbone conformational sampling of BPTI observed in a 10-μs simulation. The Φ/Ψ angles observed in the crystal structure (PDB code 5PTI)[127] are shown as gray points. Overall retention of the crystal conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

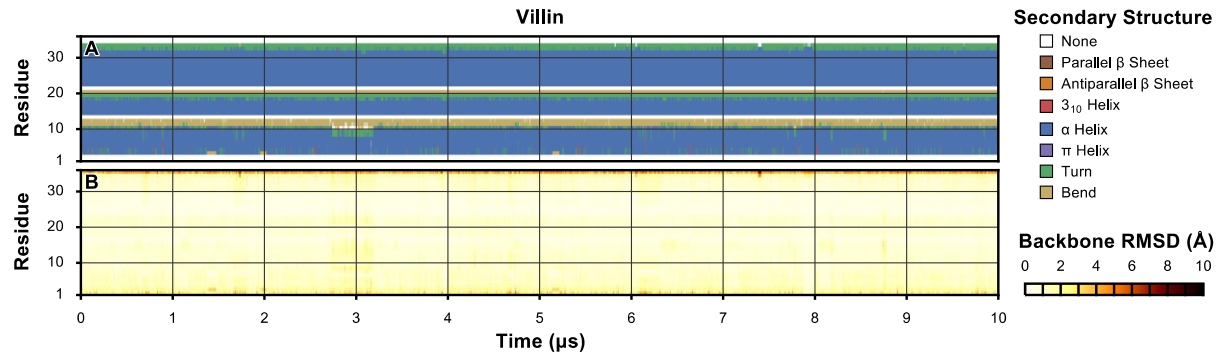

Figure 3.19. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 2F4K)[131] (B) of the double-norleucine mutant of the villin headpiece subdomain observed over the course of a 10-μs simulation.

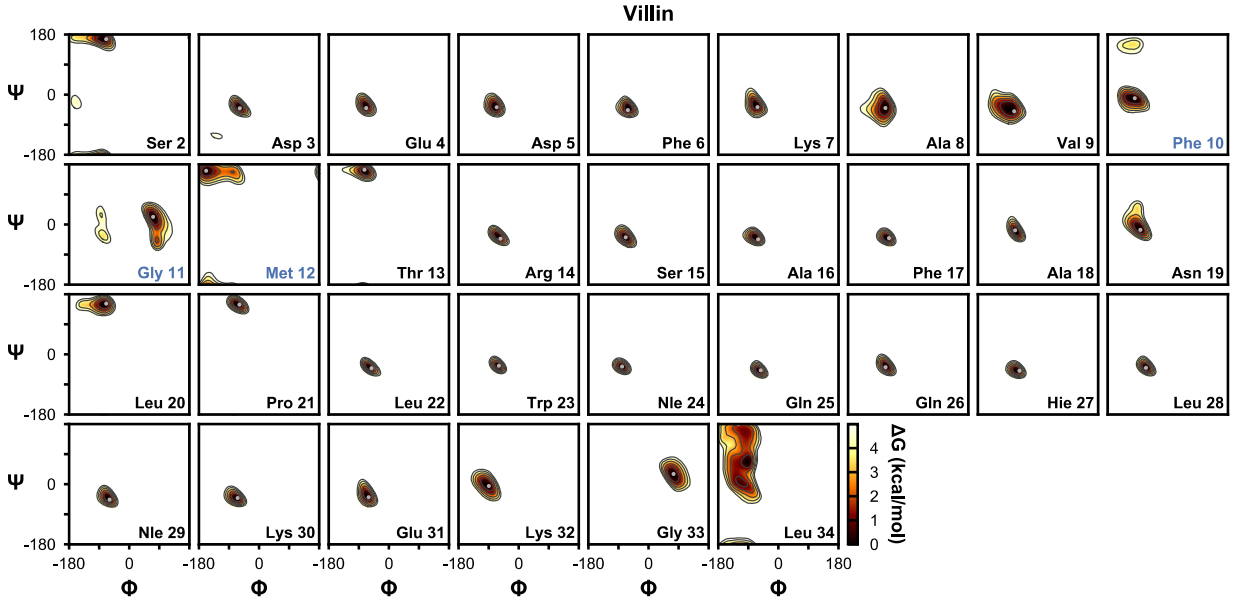

Figure 3.20. Backbone conformational sampling of the double-norleucine mutant of the villin headpiece subdomain observed in a 10-μs simulation. The Φ/Ψ angles observed in the crystal structure (PDB code 2F4K)[131] are shown as gray points. Overall retention of the crystal conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

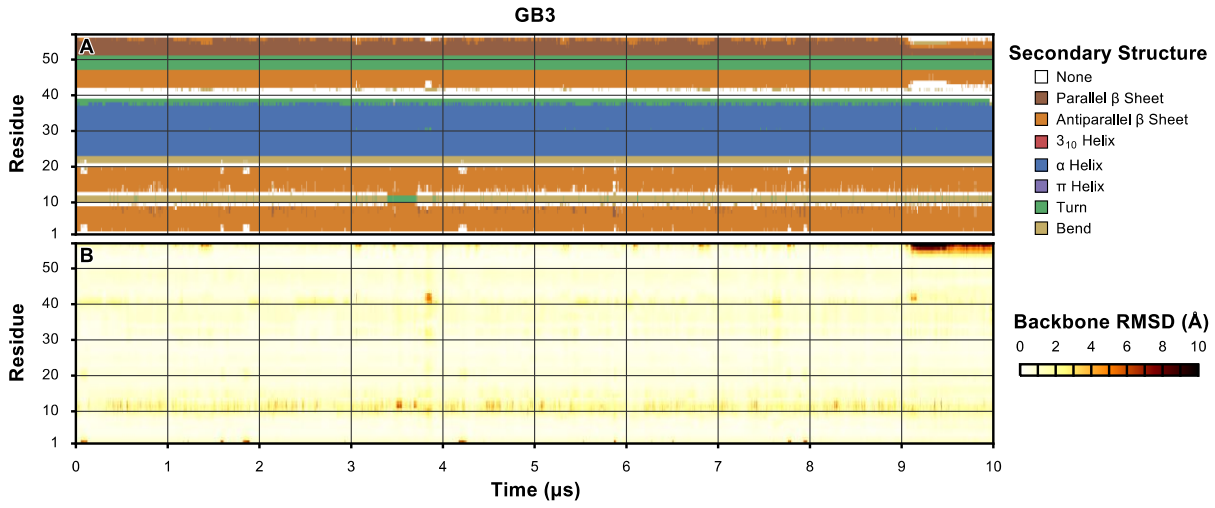

Figure 3.21. Secondary structure (A) and per-residue backbone RMSD relative to the NMR structure (PDB code 1P7E)[128] (B) of GB3 observed over the course of a 10-μs simulation.

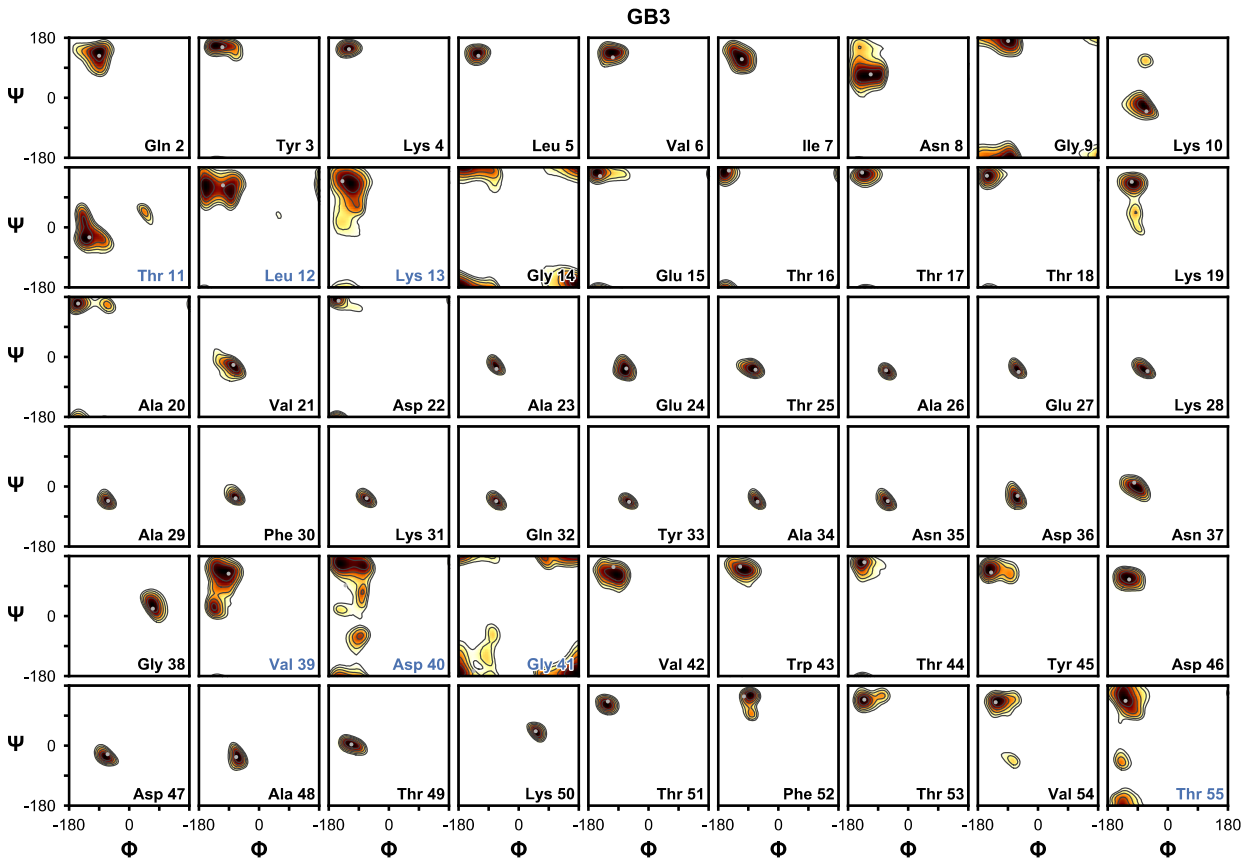

Figure 3.22. Backbone conformational sampling of GB3 observed in a 10-μs simulation. The Φ/Ψ angles observed in the NMR structure (PDB code 1P7E)[128] are shown as gray points. Overall retention of the NMR conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

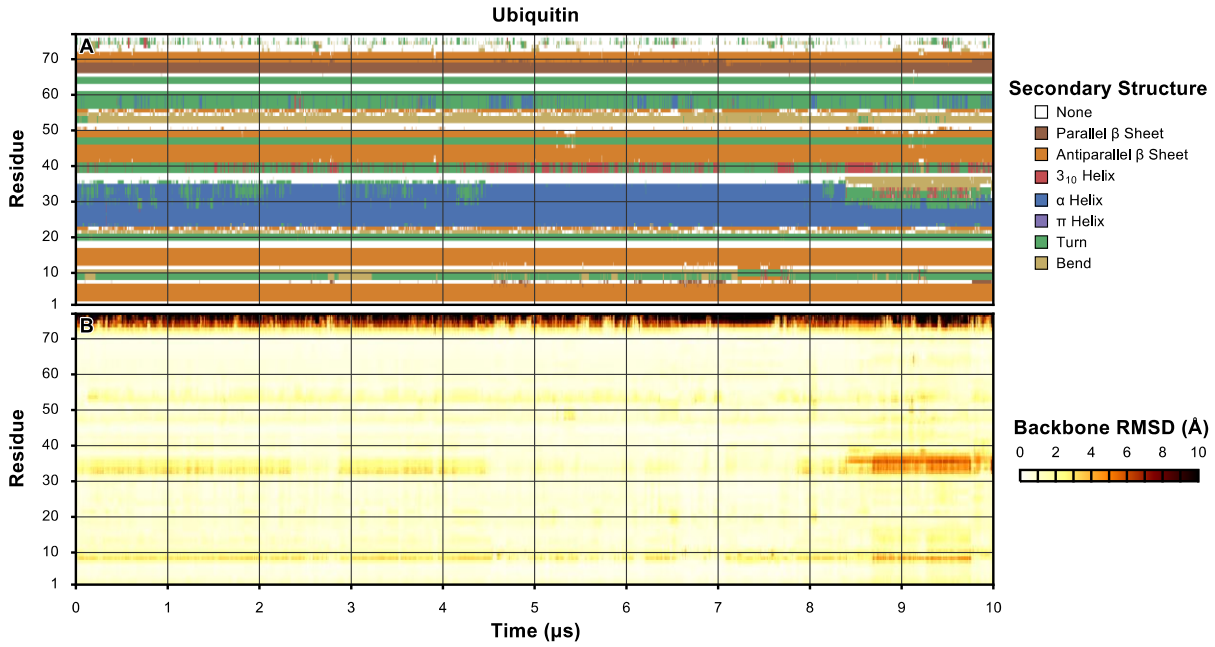

Figure 3.23. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 1UBQ)[130] (B) of ubiquitin observed over the course of a 10-μs simulation.

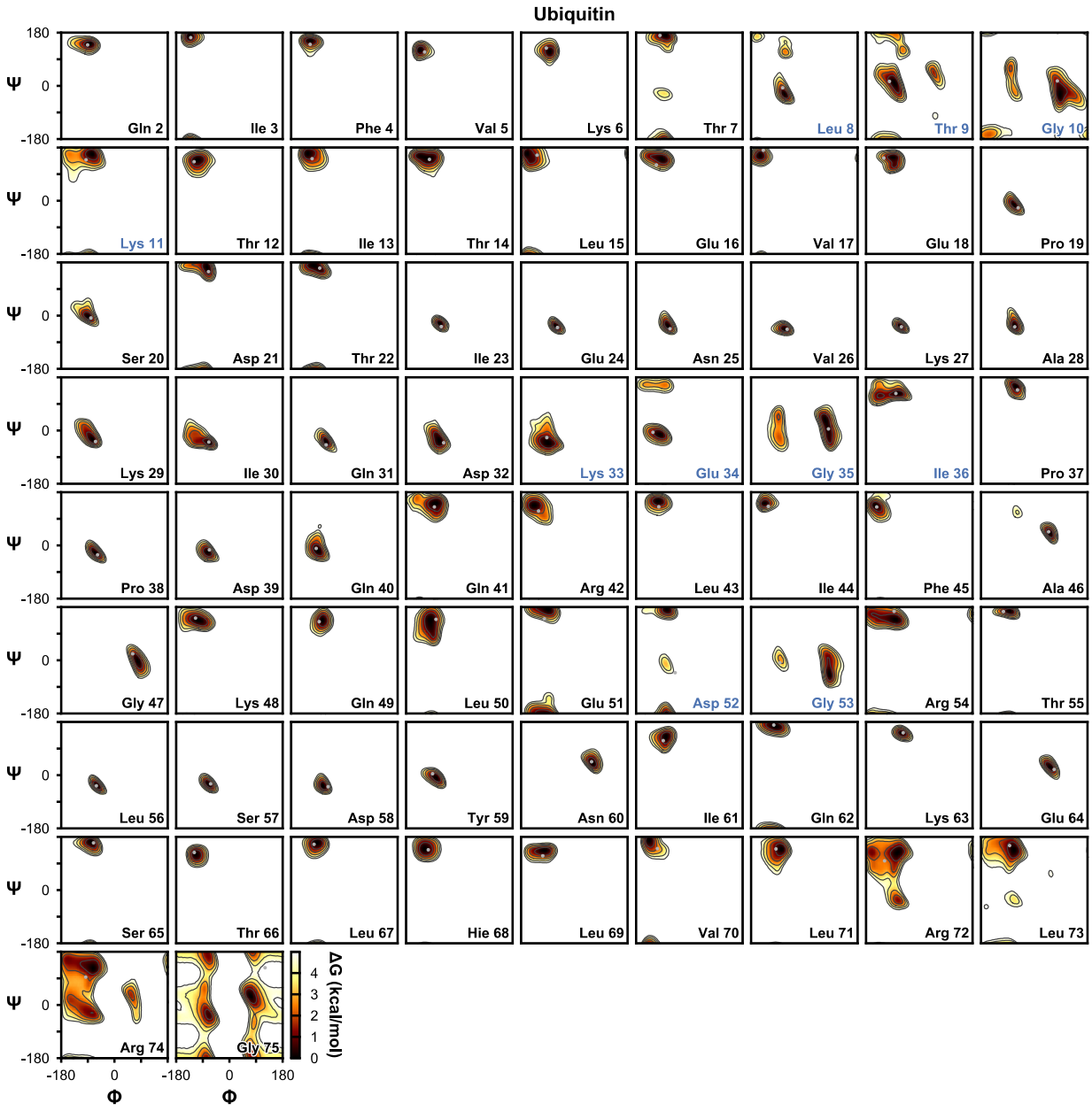

Figure 3.24. Backbone conformational sampling of ubiquitin observed in a 10-μs simulation. The Φ/Ψ angles observed in the crystal structure (PDB code 1UBQ)[130] are shown as gray points. Overall retention of the crystal conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

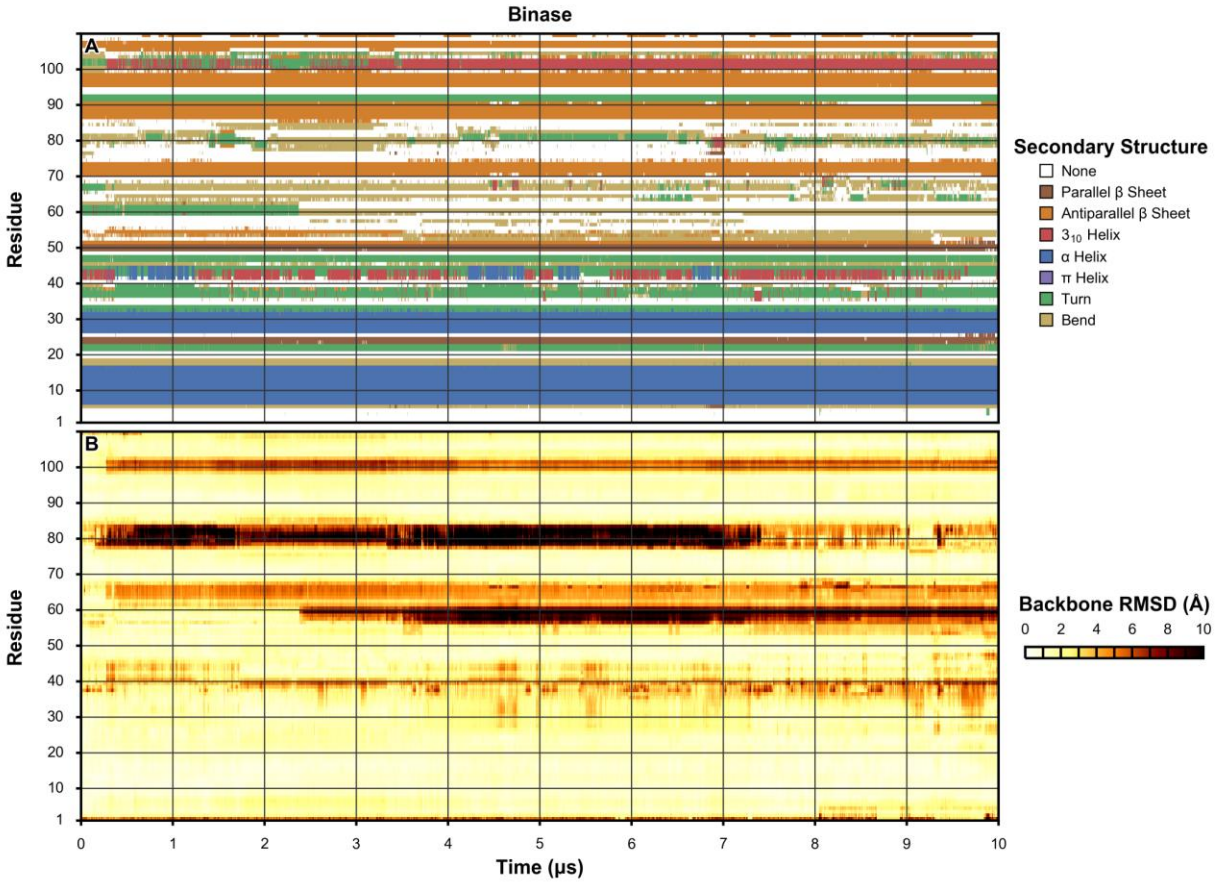

Figure 3.25. Secondary structure (A) and average per-residue backbone RMSD relative to the twenty structures of the NMR ensemble (PDB code 1BUJ)[126] (B) of binase observed over the course of a 10-μs simulation.

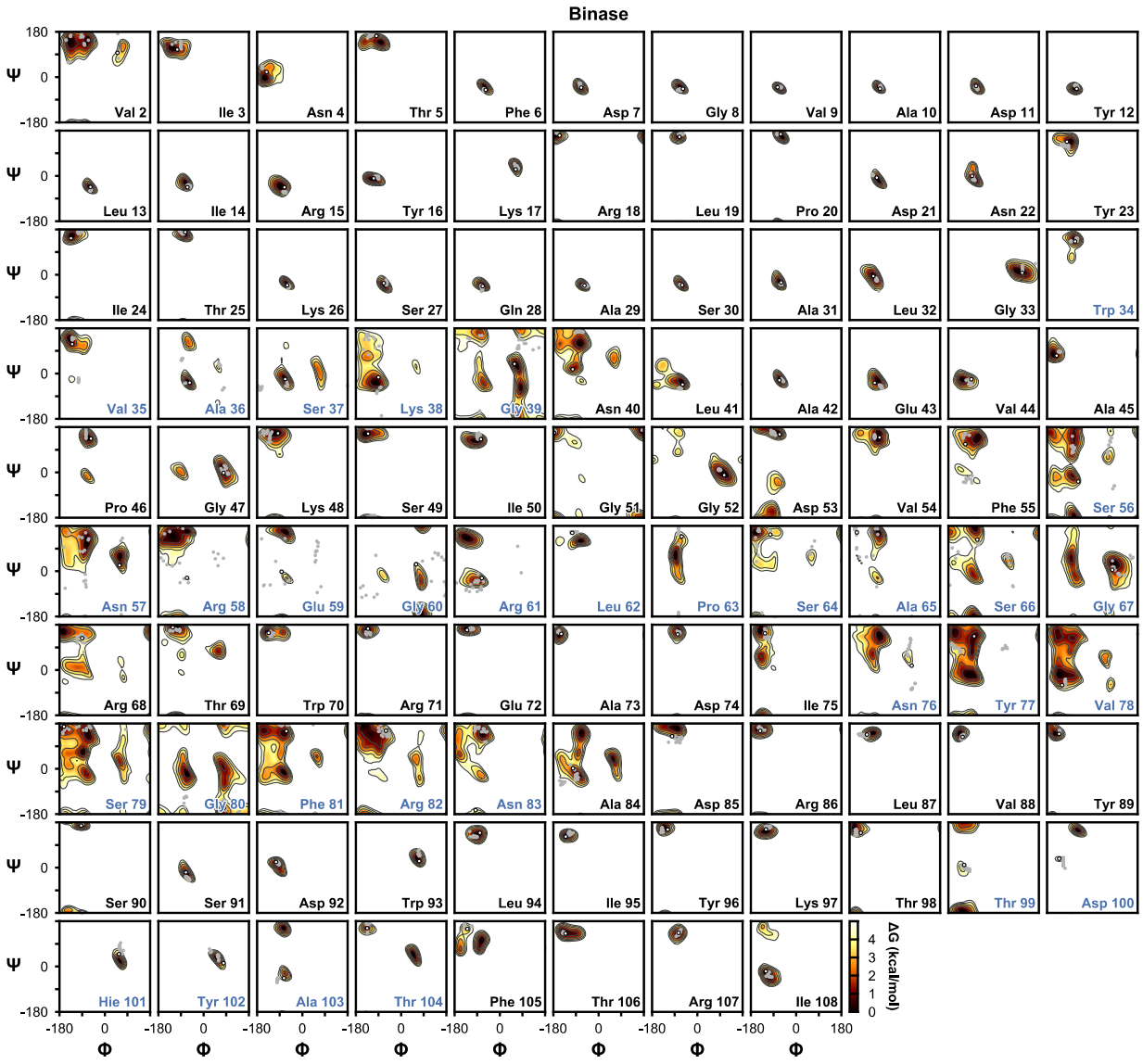

Figure 3.26. Backbone conformational sampling of binase observed in a 10-μs simulation. The Φ/Ψ angles observed in the twenty structures of the NMR ensemble (PDB code 1BUJ)[126] are shown as gray points, while those observed in the crystal structure (PDB code 1GOU)[157] are shown as white points. Overall retention of the crystal conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

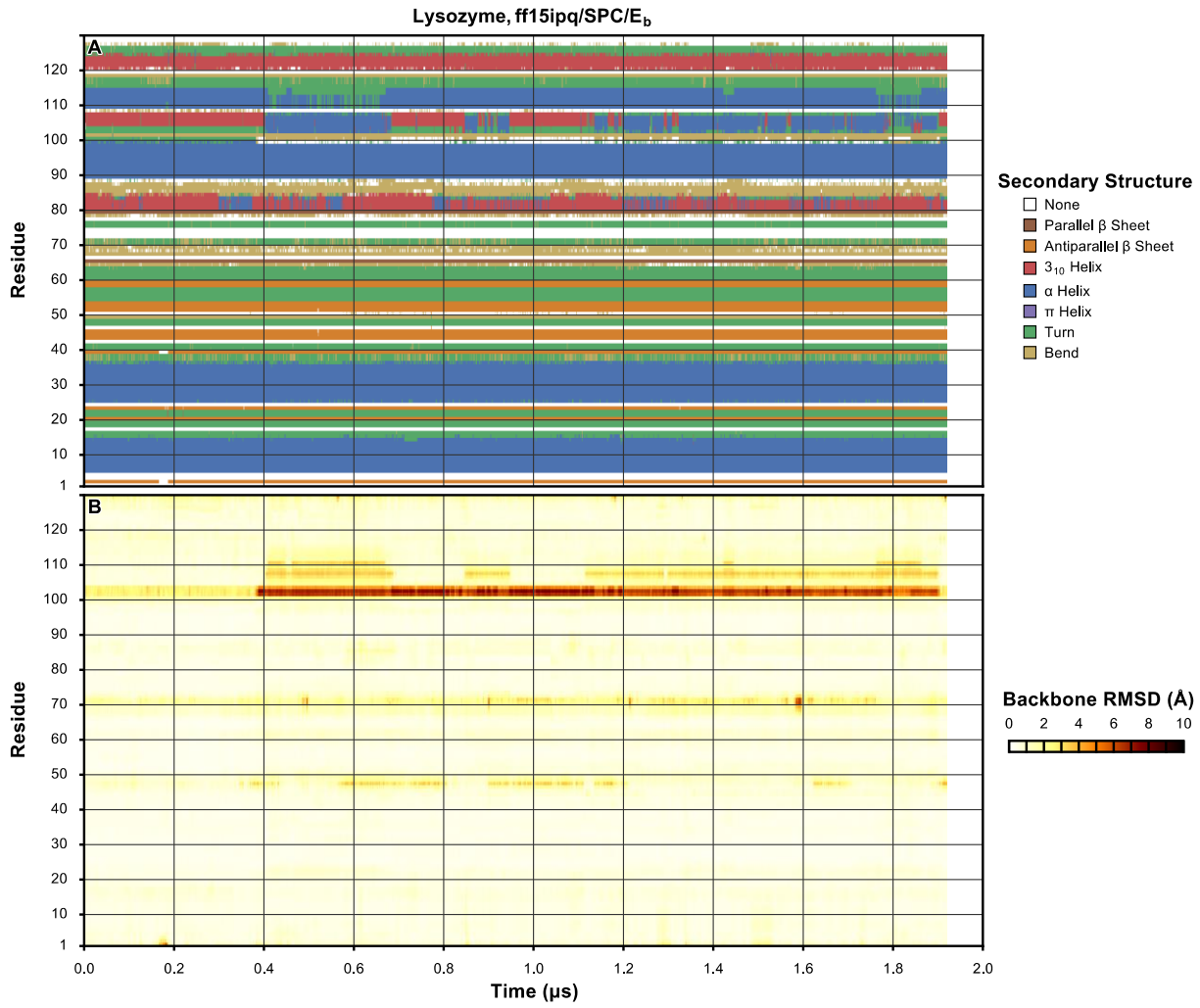

Figure 3.27. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 4LZT)[129] (B) of lysozyme observed over the course of a 2-μs simulation.

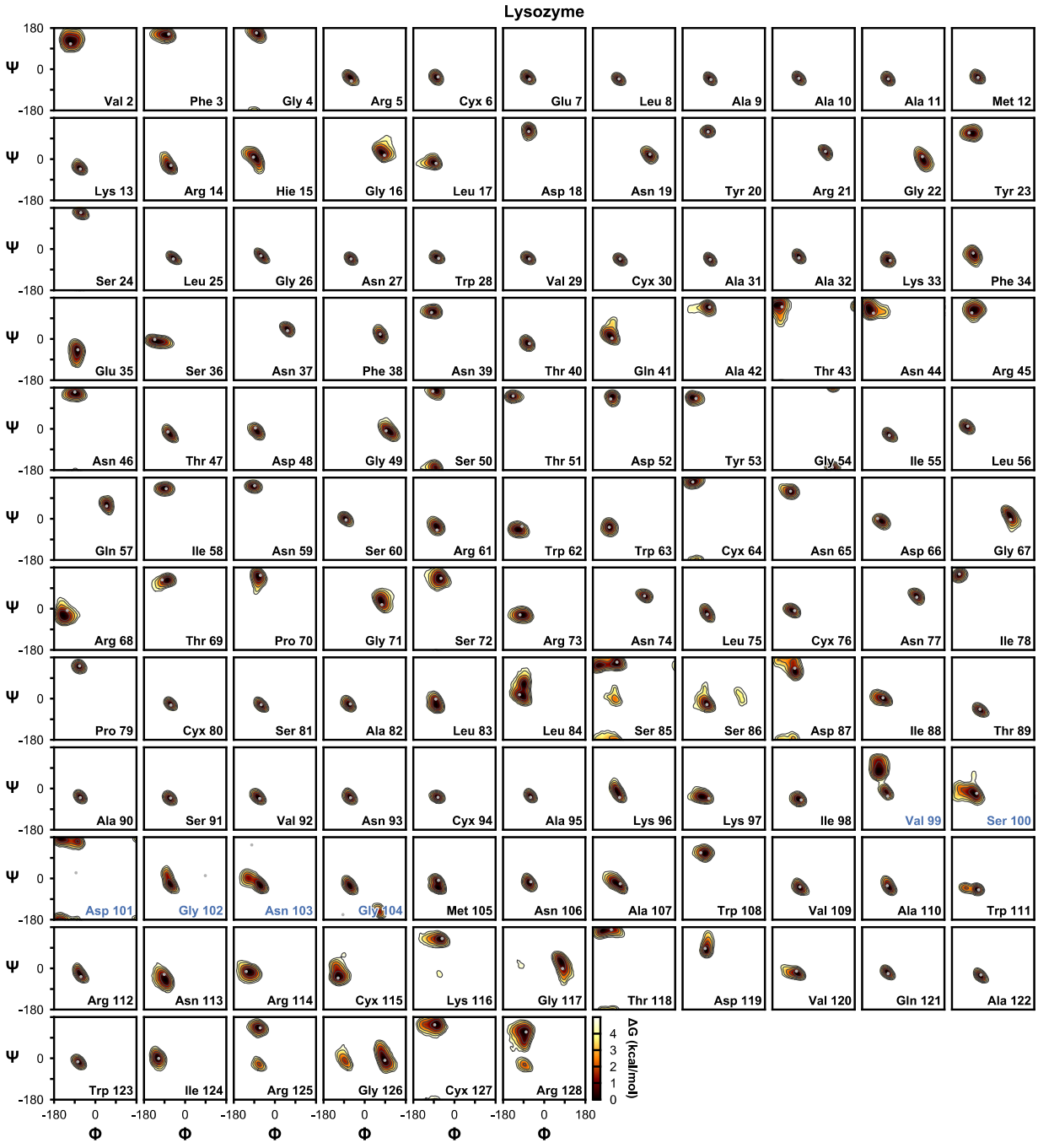

Figure 3.28. Backbone conformational sampling of lysozyme observed in a 10-μs simulation. The Φ/Ψ angles observed in the crystal structure (PDB code 4LZT)[129] are shown as gray points. Overall retention of the crystal conformation of most residues is good; regions of deviation mentioned in the main text are highlighted in blue.

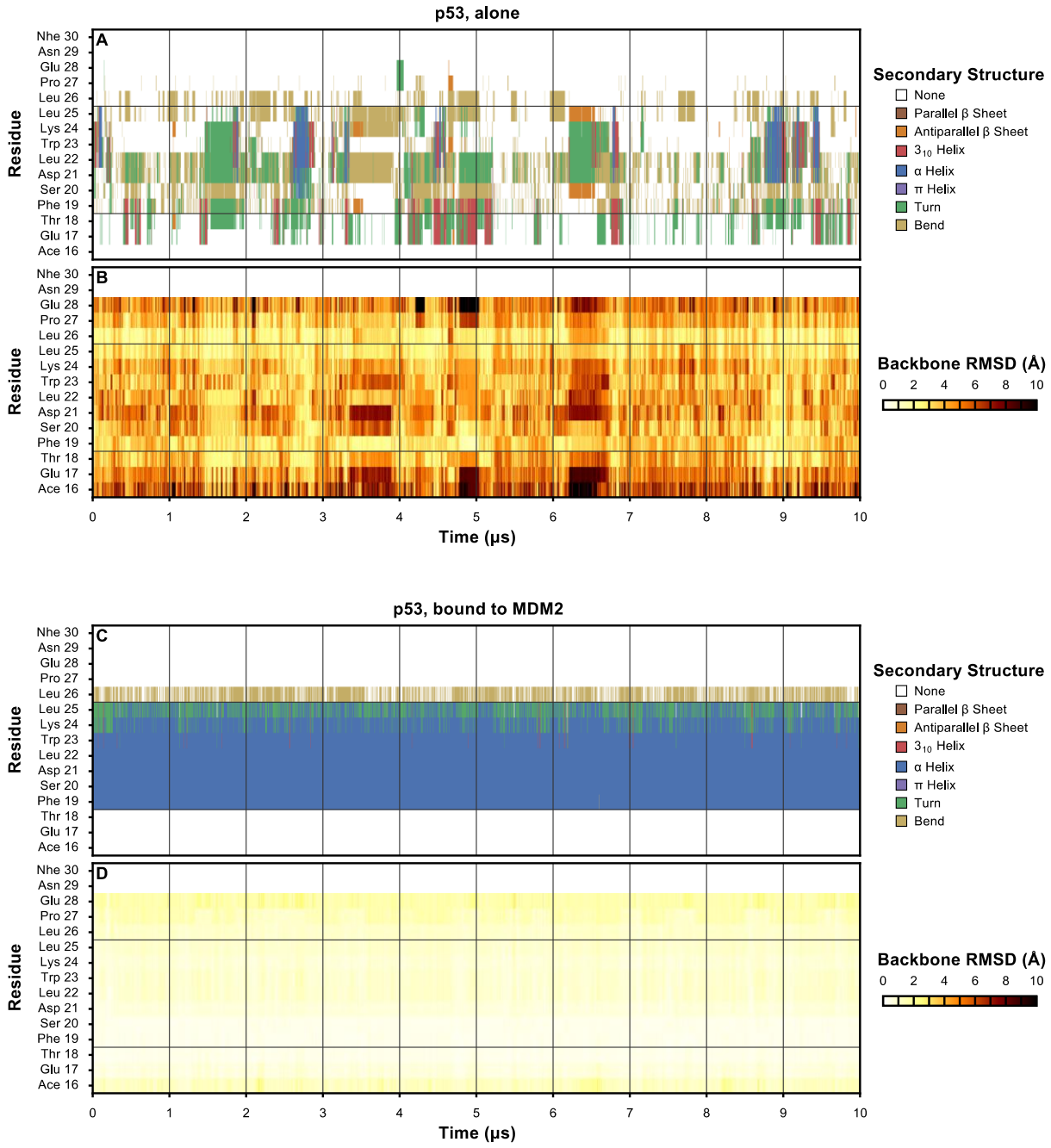

Figure 3.29. Secondary structure (A, C) and per-residue backbone RMSD relative to the crystal structure (PDB code 1YCR)[132] (B, D) of p53 observed over the course of 10-μs simulations alone (A, B) and in complex with MDM2 (C, D). Horizontal gridlines indicate the portion of p53 that forms an α-helix in the experimental structure of the p53/MDM2 complex.

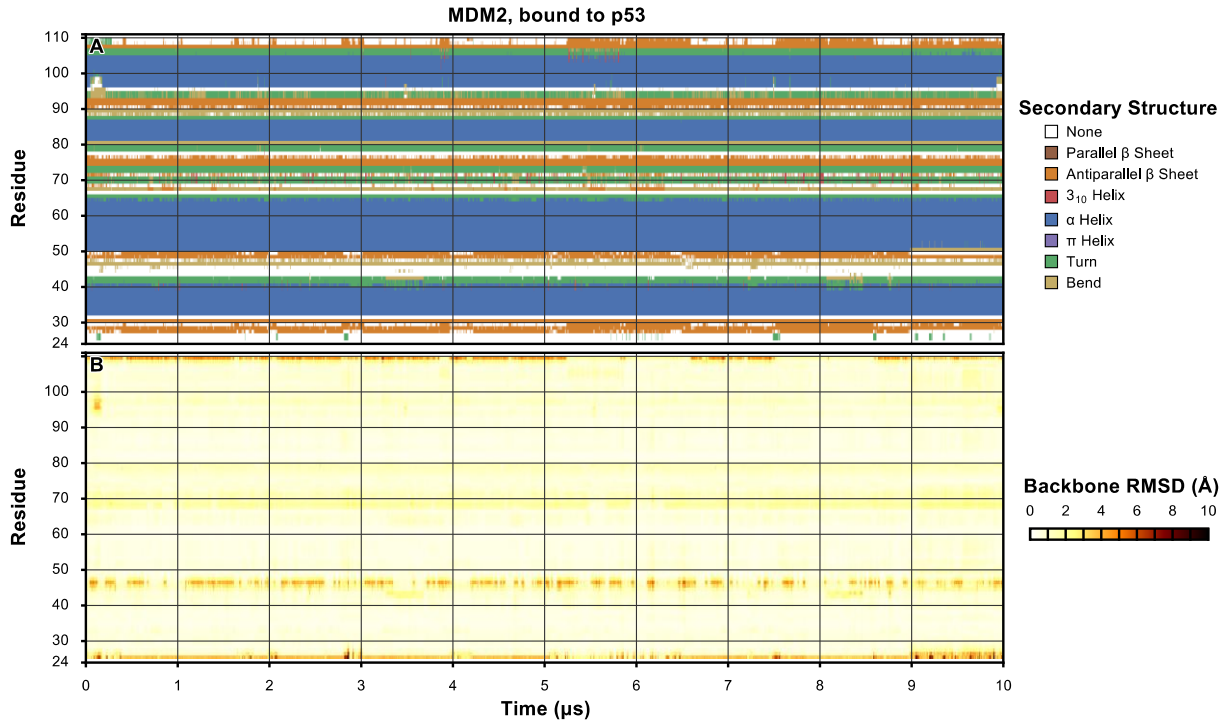

Figure 3.30. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 1YCR)[132] (B) of MDM2 observed over the course of a 10-μs simulation in complex with p53.

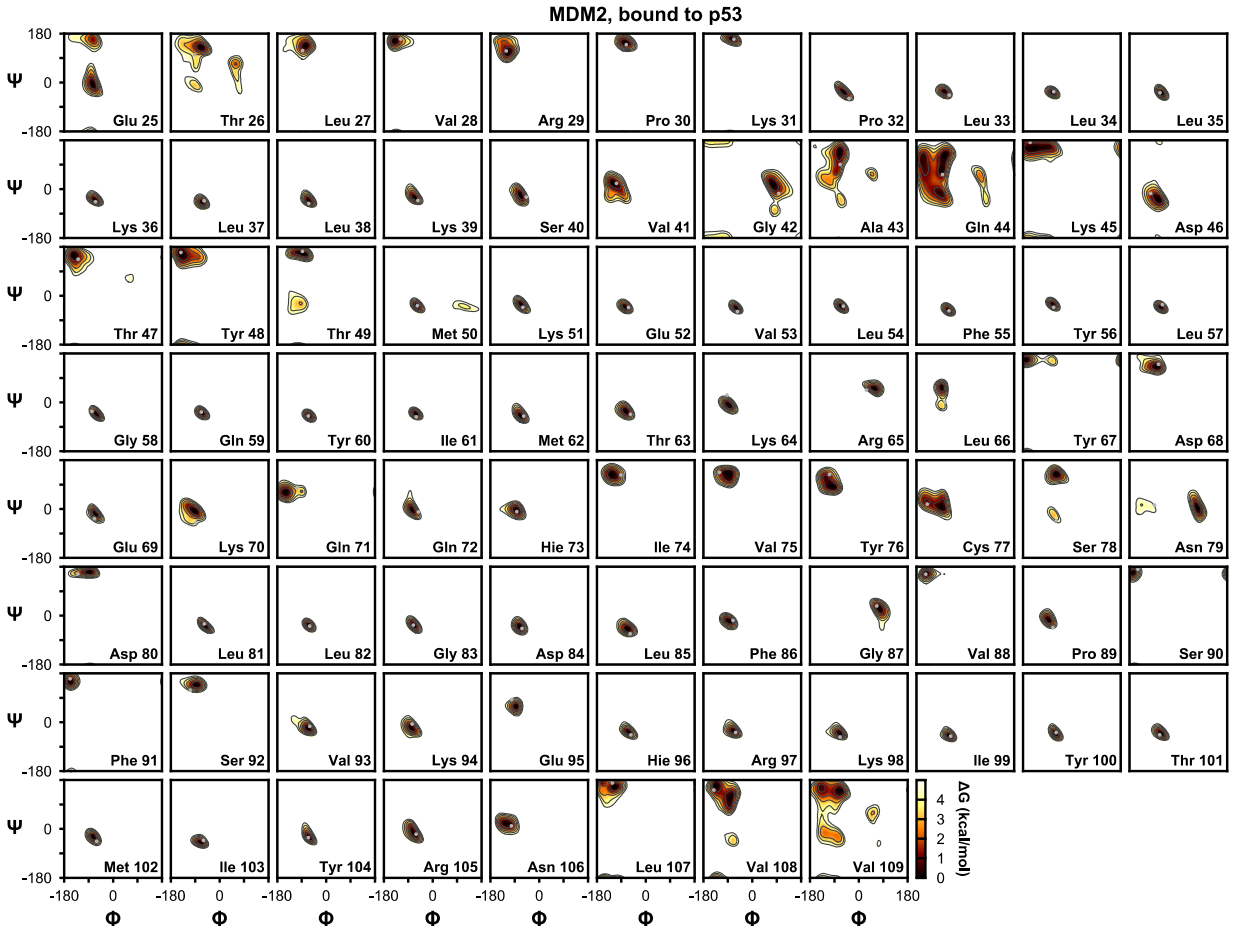

Figure 3.31. Backbone conformational sampling of MDM2 observed in a 10-μs simulation in complex with p53. The Φ/Ψ angles observed in the crystal structure (PDB code: 1YCR)[132] are shown as gray points.

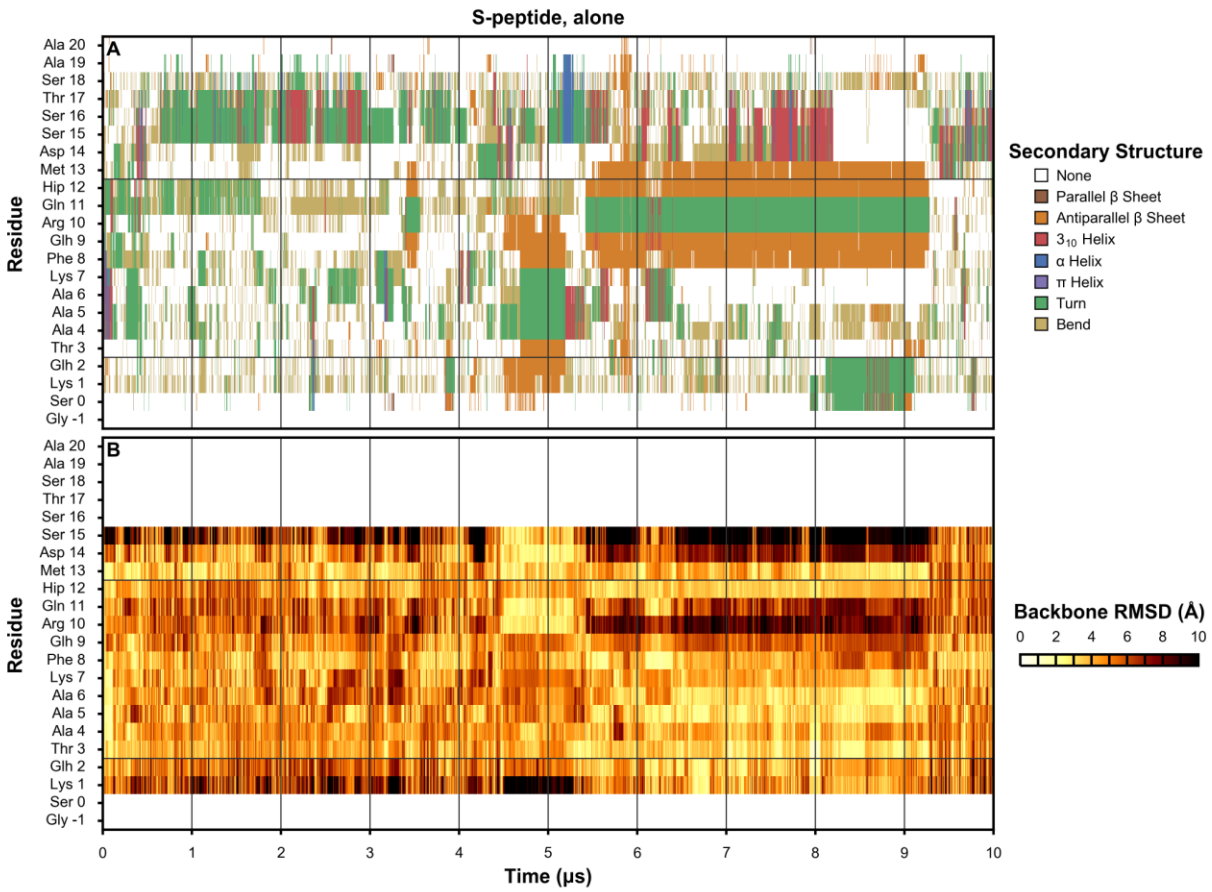

Figure 3.32. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 1RNU)[133] (B) of S-peptide observed over the course of a 10-μs simulation alone. Horizontal gridlines indicate the portion of S-peptide that forms an α-helix in the experimental structure of the S-peptide/S-protein complex.

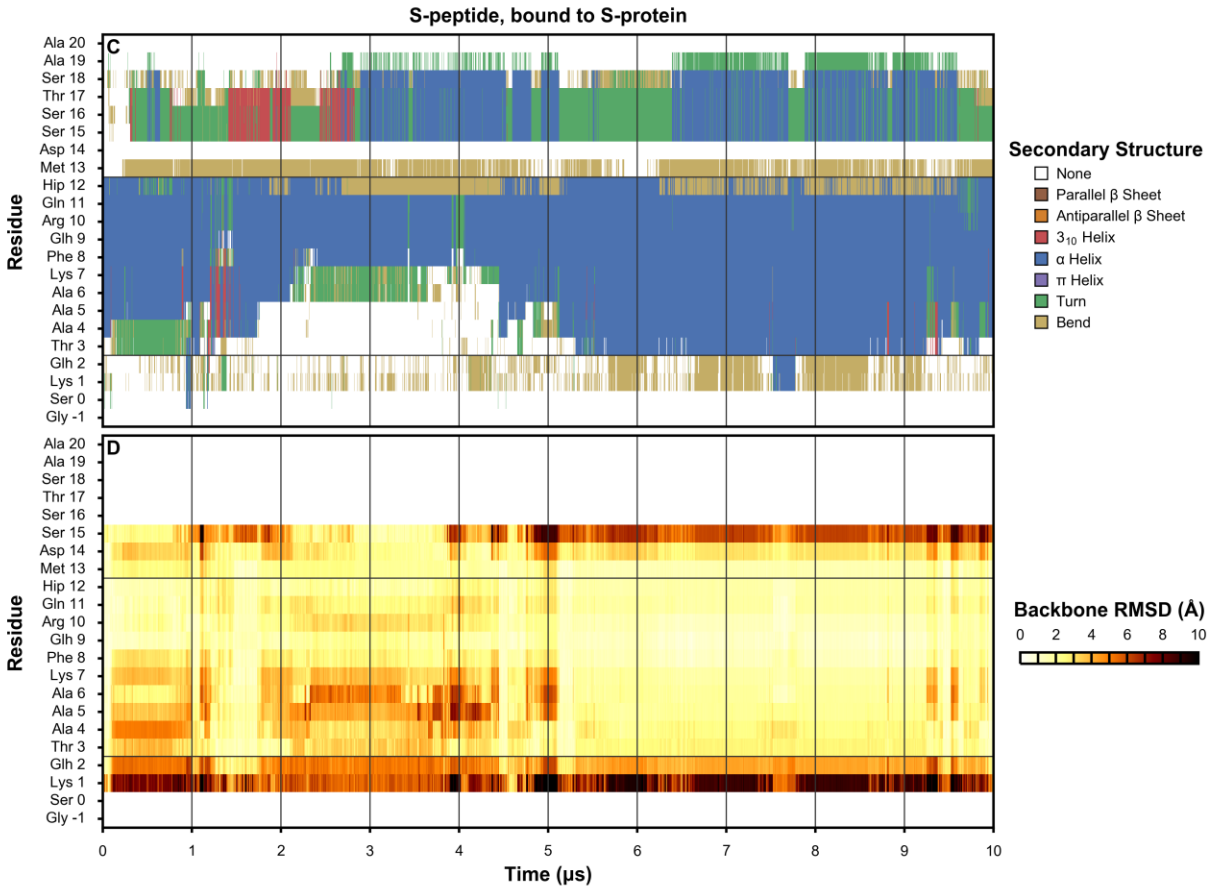

Figure 3.32 (Continued). Secondary structure (C) and per-residue backbone RMSD relative to the crystal structure (PDB code 1RNU)[133] (D) of S-peptide observed over the course of a 10-μs simulation in complex with S-protein. Horizontal gridlines indicate the portion of S-peptide that forms an α-helix in the experimental structure of the S-peptide/S-protein complex.

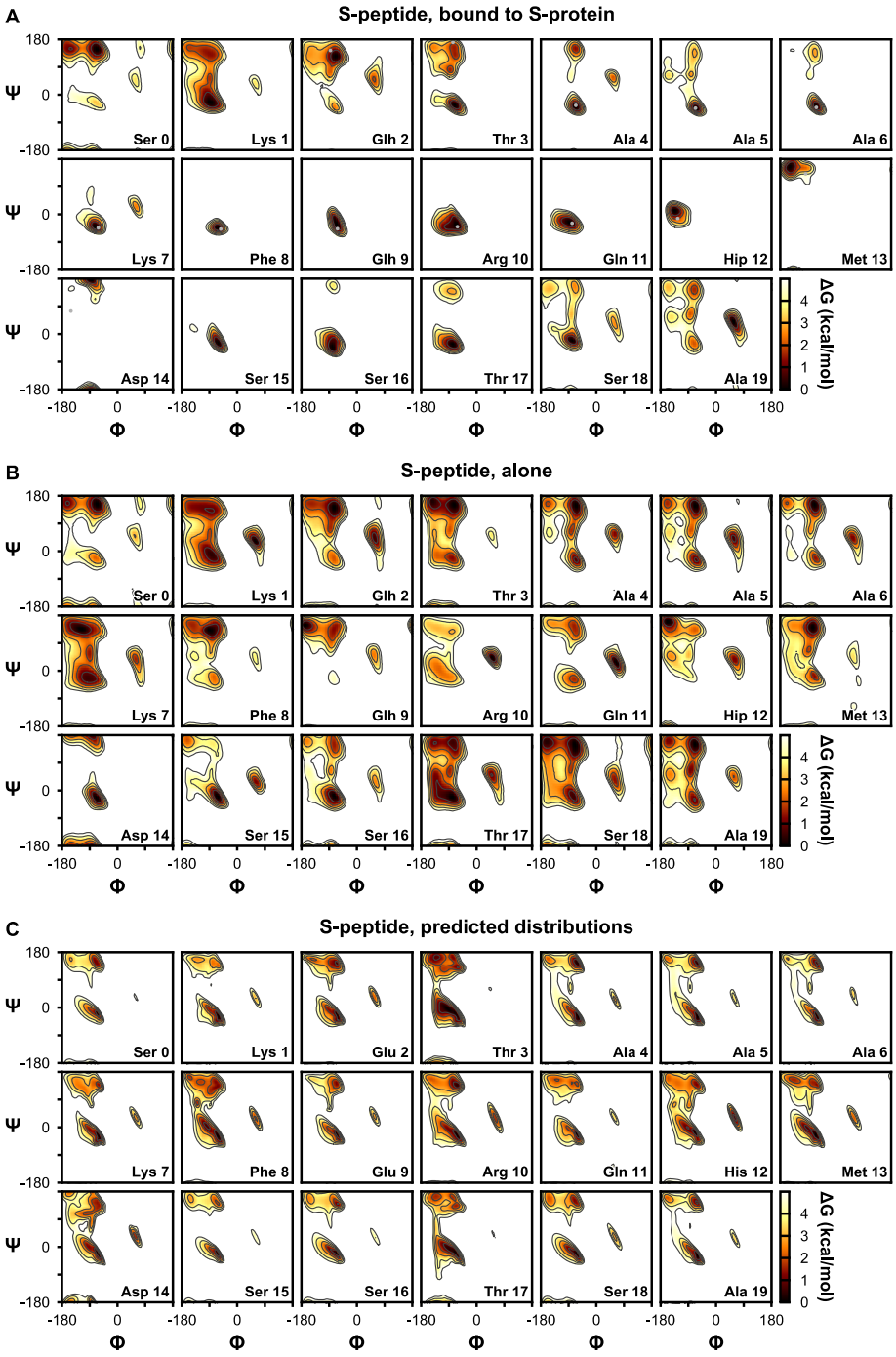

Figure 3.33 Backbone conformational sampling of S-Peptide observed in 10-μs simulations in complex with S-protein (A) and alone (B). The Φ/Ψ angles of residues present and resolved in the crystal structure of the S-peptide/S-protein complex (PDB code 1RNU)[133] are shown as gray points. For comparison are shown distributions for the S-peptide sequence obtained from the Neighbor-Dependent Ramachandran Distribution (NDRD) dataset, derived from conformations observed in the loops of solved structures (C).[143]

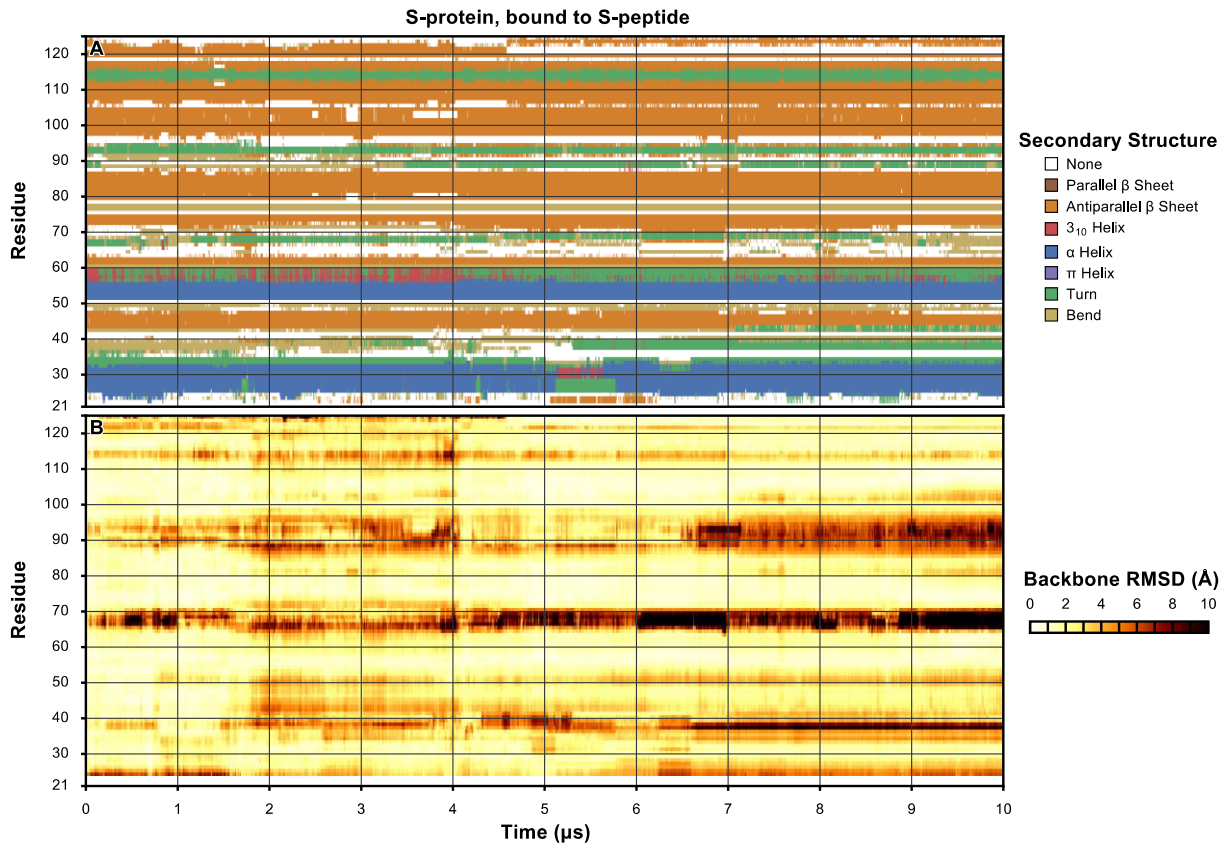

Figure 3.34. Secondary structure (A) and per-residue backbone RMSD relative to the crystal structure (PDB code 1RNU)[133] (B) of S-protein observed over the course of a 10-μs simulations in complex with S-peptide.

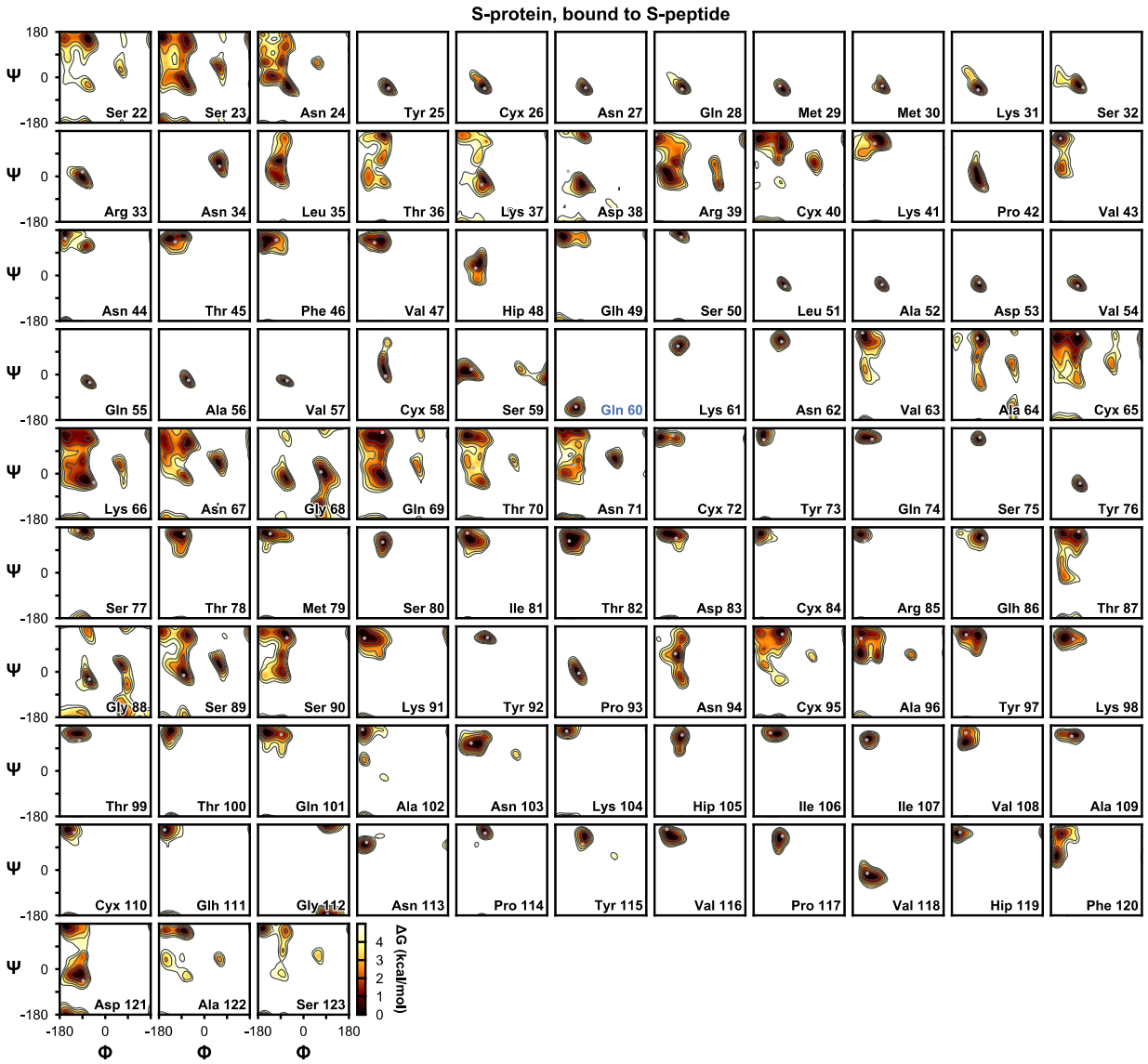

Figure 3.35. Backbone conformational sampling of S-protein observed in a 10-μs simulation in complex with S-peptide. The Φ/Ψ angles observed in the crystal structure (PDB code: 1RNU)[133] are shown as gray points. Overall retention of the crystal conformation of most residues is good; Gln 60, which retains its uncommon 'plateau' conformation throughout the simulation, is highlighted in blue.

3.7.3 Supporting tables

Table 3.4. Backbone torsion classes and terms applied to heavy atoms for 28 nonterminal residue forms.

Residue	Φ/Ψ Class	Φ'/Ψ' Subclass	φ	ψ	φ'	ψ'
Ala	Neutral	Alanine	C-N-CX-C	N-CX-C-N	C-N-CX-CT	CT-CX-C-N
Ash	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Asn	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Cys	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Cyx	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Glh	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Gln	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Leu	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Lyn	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Met	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Nle	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Ser	Neutral	Two-branched	C-N-CX-C	N-CX-C-N	C-N-CX-2C	2C-CX-C-N
Ile	Neutral	Three-branched	C-N-CX-C	N-CX-C-N	C-N-CX-3C	C-N-CX-3C
Thr	Neutral	Three-branched	C-N-CX-C	N-CX-C-N	C-N-CX-3C	C-N-CX-3C
Val	Neutral	Three-branched	C-N-CX-C	N-CX-C-N	C-N-CX-3C	C-N-CX-3C
Hid	Neutral	Aromatic	C-N-CX-C	N-CX-C-N	C-N-CX-TA	TA-CX-C-N
Hie	Neutral	Aromatic	C-N-CX-C	N-CX-C-N	C-N-CX-TA	TA-CX-C-N
Phe	Neutral	Aromatic	C-N-CX-C	N-CX-C-N	C-N-CX-TA	TA-CX-C-N
Tyr	Neutral	Aromatic	C-N-CX-C	N-CX-C-N	C-N-CX-TA	TA-CX-C-N
Trp	Neutral	Aromatic	C-N-CX-C	N-CX-C-N	C-N-CX-TA	TA-CX-C-N
Asp	Negatively-Charged		C-N-TM-C	N-TM-C-N	C-N-TM-2C	2C-TM-C-N
Cym	Negatively-Charged		C-N-TM-C	N-TM-C-N	C-N-TM-2C	2C-TM-C-N
Glu	Negatively-Charged		C-N-TM-C	N-TM-C-N	C-N-TM-2C	2C-TM-C-N
Arg	Positively-Charged		C-N-TP-C	N-TP-C-N	C-N-TP-C8	C8-TP-C-N
Hip	Positively-Charged		C-N-TP-C	N-TP-C-N	C-N-TP-C8	C8-TP-C-N
Lys	Positively-Charged		C-N-TP-C	N-TP-C-N	C-N-TP-C8	C8-TP-C-N
Gly	Glycine		C-N-TG-C	N-TG-C-N		
Pro	Proline		C-TN-TJ-C	TN-TJ-C-N	C-TN-TJ-CT	CT-TJ-C-N

3.8 SUBSEQUENT DEVELOPMENTS

Having developed an improved force field that yielded more accurate stability of salt bridge interactions, and validated the force field's accuracy on globular proteins and disordered peptides, I began simulating the more complex MoCVNH3 protein with AMBER ff15ipq and the SPC/E_b water model with which it was developed. I also planned to test the CHARMM22* force field, which includes a similar correction to its salt bridge interactions, but needed to select an appropriate water model with which to pair it. It had recently been found that fixed-charge water models generally overstabilize protein-protein relative to protein-solvent interactions, and this

overstabilization had been addressed in the recently-developed TIP4P-D water model.[102] Since the CHARMM22*/TIP4P-D pair had been tested extensively during TIP4P-D's development, I elected to use this combination for my second set of simulations of MoCVNH3, as described in the following chapter.

4.0 BIOPHYSICAL CHARACTERIZATION OF THE MOCVNH3 PROTEIN

This chapter is based on a research article submitted for publication as: Debiec, K. T.; Whitley, M. J.; Koharudin, L. M. I.; Chong, L. T.; and Gronenborn, A. M. Merging Structure and Dynamics from Biophysical Experiments and Atomistic Simulations of a Model Two-Domain Protein.

4.1 INTRODUCTION

Multi-domain proteins in which the connected domains each fold and function independently are prevalent in nature.[1,2] Such proteins, through spatial and temporal coordination of their varied functional units, are capable of executing specific and tailored activities in catalysis, signaling, regulation of gene expression, and other cellular processes.[3] The individual domains are connected by inter-domain linkers whose length and composition enable them to adopt orientations that have evolved for specific biological activities and functions.[4,5] In many cases, the linkers are highly flexible, allowing the domains to adopt numerous inter-domain orientations, from which the selection of functional competent conformations may occur.[6] While most multi-domain proteins are linked linearly in sequence, roughly one tenth possess domain insertions where a 'guest' domain is implanted into a loop of a 'host' domain, such that the two domains are connected by a pair of inter-domain linkers.[7]

Characterization of the relative domain orientations within multi-domain proteins has been challenging by traditional structural biology techniques, such as X-ray crystallography, due to the inherent flexibility of inter-domain linkers, lack of density for certain segments of the polypeptide

chain, and influence of crystal packing on the positioning of domains. On the other hand, multi-domain proteins represent intriguing targets for integrative structural biology approaches, which combine results from experiments and computer simulations[5,8] by either (i) computationally generating a large ensemble of potential structural models and subsequently filtering the models based on agreement with the experimental data, or (ii) explicitly biasing the generation of structural models in accord with the experimental data. Such approaches have been particularly useful for studying flexibly linked multi-domain proteins and protein complexes,[9–15] often integrating data from NMR, SAXS, X-ray crystallography, and other experimental techniques into a single structural model. Critically, the validity of any approaches aimed at bridging the gaps between experimentally accessible and computationally generated data depends on the accuracy of the biomolecular force fields used in the computations, which dictate sampling of the conformational space for the entire system.

Traditionally, force fields have been parameterized to reproduce the properties of small molecules, with parameters derived from experiment and quantum mechanical calculations, and their accuracy for biomolecules are validated using simulations of well-characterized benchmark systems. Such systems have included small globular proteins (*e.g.* ubiquitin, GB3, and lysozyme),[81,87,92,184] although more recently, simulations of significantly more flexible, disordered peptides and proteins (*e.g.* the MDM2-binding p53 peptide and α-synuclein) have been carried out.[102,184,185] The latter have revealed that most force fields, when paired with their intended explicit water models, suffer from an imbalance between protein-protein and protein-water interactions, yielding conformations of nonglobular proteins that are much more compact than experimentally observed, as well as overstabilizing the folded states of globular proteins.[102] Since the conformational space accessible to nonglobular systems is very large, exhaustive sampling is

beyond the capabilities of current simulation methods, and experimentally well-characterized systems are necessary to serve as a link between globular and nonglobular proteins. In particular, simultaneous validation of both the interior protein structure and the balance between protein-protein and protein-water interactions is needed. Flexibly linked, multi-domain proteins present an ideal opportunity to fulfill this requirement, since the inter-domain conformational space of such proteins is large, yet sufficiently restricted to be addressed with current simulation methods. Such affordable, yet complex model systems are becoming increasingly valuable as computational methods shift towards more intricate and expensive algorithms, such as implemented in the AMOEBA and CHARMM Drude polarizable force fields.[107,108]

An ideal test system among flexibly linked multi-domain proteins is the relatively small, two-domain protein MoCVNH3 that has been structurally characterized by our group using both NMR spectroscopy and X-ray crystallography.[16,17] MoCVNH3 is a domain-insertion protein in which a 'guest' LysM domain is inserted into a surface loop of a 'host' Cyanovirin-N Homology (CVNH) domain, positioning the LysM domain between the two pseudo-symmetric halves of two-lobed CVNH domain.[18] This protein is found in *Magnaporte oryzae*, an ascomycete fungus that causes rice blast disease, the most devastating infection of cultivated rice, which destroys crops in unprecedented amounts worldwide.[19] Functionally, both CVNH and LysM are carbohydrate-binding domains: CVNH binds to mannose sugars, while LysM interacts with GlcNAc-containing carbohydrates such as peptidoglycan and chitin.[20,21] The binding of carbohydrates by each domain in MoCVNH3 is independent of the other, with no communication between the domains.[22] While the wild-type protein could not be crystallized, complete removal of the inter-domain linkers yielded a construct that crystallized and maintained the ability to bind both carbohydrate ligands. A comparison of the resulting crystal structure with the NMR structure of wild-type MoCVNH3

revealed that the absence of the linkers has no effect on the structures of the individual domains.[17] However, although the domain structures of wild-type MoCVNH3 were solved to high resolution by NMR, no fixed relative domain orientations were compatible with the solution data, due to the lack of inter-domain restraints.[22]

Here, we investigated the influence of inter-domain linker length on the overall structure and dynamics of MoCVNH3, as well as the conformational space of inter-domain orientations in solution, using an integrated approach that combines biophysical experiments and molecular dynamics (MD) simulations. In particular, we performed SAXS, NMR relaxation, and paramagnetic relaxation enhancement (PRE) experiments along with microsecond (μs)-scale MD simulations in explicit solvent. In carrying out the simulations, we compared the accuracy of two biomolecular force field/water model combinations in modeling the structure and dynamics of this tethered two-domain protein, and establish that each combination is accurate for certain properties and inaccurate for others. Overall, we demonstrate that an integrated approach, incorporating different experimental and computational methods, permits characterization of both the inter-domain orientations and dynamic of multi-domain proteins, using the MoCVNH3 protein as an example.

4.2 MATERIALS AND METHODS

4.2.1 Protein expression and purification

Proteins were expressed and purified as described previously for wild-type MoCVNH3.[22] In brief, pET-15b(+) vectors containing the different coding sequences for the individual protein constructs

were used to transform *E. coli* Rosetta2 (DE3) cells (Novagen). All constructs encoded N-terminal polyhistidine tags, followed by a TEV protease cleavage site. After cleavage, the native protein N-terminus was obtained, removing a four-residue addition that was present in the earlier protein constructs. Mutant coding sequences were created using the QuikChange XL II site-directed mutagenesis kit (Stratagene). Cells were initially grown at 37 °C, induced with 1 mM isopropyl β-D-1-thiogalactopyranoside at an OD_{600} of ~0.8, and further grown for 18 hr at 16 °C for protein expression. Cells were harvested by centrifugation, resuspended in lysis buffer (25 mM Tris-HCl (pH 8.0), 150 mM NaCl, 5 mM DTT, and 3 mM NaN_3) and ruptured by passage through a microfluidizer (MicroFluidics M-110Y, Hyland Scientific). Cell debris was removed by ultracentrifugation (19,000 RPM), and the supernatant was loaded onto an Ni^{2+}-derivatized HisTrap column (GE Healthcare), pre-equilibrated with loading buffer (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM DTT, and 3 mM NaN_3). Proteins were eluted using a linear (25–500 mM) imidazole gradient in the same buffer and protein-containing fractions were subjected to TEV digestion in 25 mM Tris-HCl buffer, pH 8.0, 25 mM NaCl, 5 mM DTT, and 3 mM NaN_3 for removal of the N-terminal polyhistidine tag. Further purification of the cleaved proteins involved gel filtration on a Superdex75 column (GE Healthcare) in 25 mM sodium acetate buffer, pH 5.0, 25 mM NaCl, 5 mM DTT, and 3 mM NaN_3, and cation exchange on an HiTrap SP column (GE Healthcare), pre-equilibrated with loading buffer (25 mM sodium acetate (pH 5.0), 25 mM NaCl, 5 mM DTT, 3 mM NaN_3), and elution by a linear (0-250 mM) NaCl gradient in the same buffer. Protein-containing fractions were collected, buffer exchanged into 25 mM sodium acetate, pH 5.0, 25 mM NaCl, 5 mM DTT, 3 mM NaN_3, and concentrated using Centriprep devices (Millipore). For ^{15}N isotopic labeling, the bacterial cell culture was grown in modified minimal medium, containing $^{15}NH_4Cl$ as the sole nitrogen source, while for ^{13}C isotopic labeling ^{13}C-glucose was

provided as the sole carbon source. The purity and identity of all proteins were confirmed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS–PAGE) and mass spectrometry.

4.2.2 Site-selective spin-labeling

The purified single-cysteine variant of Mo-WT, Mo-SAVC (C15S, C25A, C82V), was split into two equal portions for parallel spin-labeling at the C-terminal C167 with (1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)methanethiosulfonate (MTSL) and the diamagnetic analog of MTSL: (1-acetyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)methanethiosulfonate (dMTSL, Toronto Research Chemicals Inc.), in which the oxygen on the nitroxide of MTSL is replaced with an acetyl group. Both reagents were added from 0.2 mM stocks in 10-fold molar excess to a 65 μM protein solution in 25 mM NaPhosphate buffer, pH 7.0, 25 mM NaCl, 3 mM NaN_3, and the reaction mixture was incubated at 4 °C for 16 h in the dark. Excess free tags were removed by serial dialysis against 25 mM NaAc buffer, pH 5.0, 25 mM NaCl, 3 mM NaN_3. The extent of spin-labeling (>95%) was confirmed by mass spectrometry.

4.2.3 Molecular dynamics simulations

Heavy atom coordinates of the Mo-WT and Mo-0v constructs were extracted from the solution NMR structure of Mo-WT (PDB code 2L9Y)[22] and X-ray crystal structure of Mo-0v (PDB code 58CO),[17] respectively. Coordinates of the reduced linker-length constructs Mo-2G and Mo-0G were generated using the MODELLER 9.9 software package[186] based on the coordinates of the CVNH and LysM domains from the Mo-WT NMR structure. Each system was solvated in a 105 x 105 x 105 Å cubic box, which generated a minimum solute-wall distance of 22 Å for the largest

two-domain construct (Mo-WT). Amino acid side chains were assigned protonation states consistent with the experimental pH of 5.0; *i.e.* arginine, lysine, and histidine residues were protonated while aspartate and glutamate were deprotonated. The net positive charges on the protein were neutralized through the addition of Cl^- ions; additional Na^+ and Cl^- ions were added to be consistent with the experimental salt concentration of 25 mM. Identical system configurations were used for simulations run using the AMBER software package[93] and Anton[25].

Simulations run using AMBER were carried out using the GPU implementation of the *pmemd* module.[93,97,134] Prior to running production simulations, each system was subjected to energy minimization, followed by a three-stage equilibration. In the first stage, a 20-ps simulation of the energy minimized system was carried out at constant temperature, while restraining the solute heavy atoms to their initial positions using a harmonic potential with a force constant of 1 kcal/(mol $Å^2$). In the second stage, a 1-ns simulation was carried out at constant pressure with the same harmonic restraints on positions. Finally, an additional 1-ns unrestrained simulation was carried out at constant temperature and pressure. The temperature was maintained at 25 °C using a Langevin thermostat (frictional constant of 0.1 ps^{-1}) while the pressure was maintained at 1 atm using a Monte Carlo barostat (200 fs between attempts to change the system volume).[58] Van der Waals and short-range electrostatic interactions were truncated at 10 Å; long-range electrostatic interactions were calculated using the particle mesh Ewald method.[61] To enable a 4-fs time step, hydrogens were constrained to their equilibrium values using the SHAKE and SETTLE algorithms, and hydrogen mass repartitioning was used.[135–137] The masses of solute hydrogen atoms were increased by a factor of three, and that of their attached heavy atoms decreased by the corresponding amount, such that the total mass remained constant; the masses of water molecules were not repartitioned. Coordinates were saved every 100 ps for analysis.

Simulations run on the Anton special-purpose supercomputer were equilibrated using the Desmond 3.0.1.0 software package.[25,57] Each system was subjected to energy minimization followed by a 20-ps equilibration at constant temperature, and a 1-ns equilibration at constant pressure. The temperature was maintained at 25 °C and the pressure at 1 atm, using the Martyna-Tobias-Klein thermostat and barostat (time constants of 1 ps and 2 ps, respectively).[59] To enable a 2.5-fs time step, hydrogen were constrained to their equilibrium values using the M-SHAKE algorithm.[60] A short-range non-bonded cutoff of 10 Å was used, and long-range electrostatics were calculated using the particle mesh Ewald method.[61] Production simulations were carried out at constant pressure using a 512-node Anton special-purpose supercomputer and the Multigrator integrator.[25,182] The temperature was maintained at 25 °C using the Nosé-Hoover thermostat and the pressure at 1 atm using the Martyna-Tobias-Klein barostat (time constants of 1 ps).[59,62] To enable a 2.5-fs time step, bonds to hydrogen were constrained to their equilibrium values using the M-SHAKE algorithm.[60] Van der Waals and short-range electrostatic interactions were truncated at 10 Å; long-range electrostatic interactions were calculated using the Gaussian split Ewald method,[63] and were updated every third time step. Coordinates were saved every 105 ps for analysis.

Analyses of MD simulations were carried out primarily using the AmberTools *cpptraj* program.[138] Secondary structure was assigned using the DSSP method,[139] rotational correlation times (τ_c) were calculated using the method of Wong *et al.*,[29] and NMR relaxation rates were calculated using the iRED method.[140] Small-angle X-ray scattering curves were calculated using the saxs_md and CRYSOL programs,[187,188] and standard errors were estimated using a block averaging method.[64]

4.2.4 Small-angle X-ray scattering

Small-angle X-ray scattering data were collected for the Mo-WT, Mo-2G, Mo-0G, and Mo-0v constructs at 25 °C. Samples were prepared in 25 mM NaAc buffer (pH 5.0), 25 mM NaCl, 5 mM DTT, and 3 mM NaN_3. For the Mo-WT, Mo-2G, and Mo-0G constructs, data were collected using protein concentrations of 5.0, 2.5, and 1.25 mg/mL. No concentration-dependent effects were observed, and the 5.0 mg/mL data are presented here. For the less soluble Mo-0v construct, data was collected at 0.55 mg/mL. All experimental SAXS data were collected at beamline 12-ID-B of the Advanced Photon Source at Argonne National Laboratory (Lemont, IL, USA) using X-rays of energy 14 keV ($\lambda \approx 0.8856$ Å). For each measurement, 30 individual exposures of 1 second each were collected, compared to check for radiation damage, and averaged to yield the final scattering curves. Buffer scattering measurements were performed in an equivalent fashion using protein-free buffer aliquots from the final purification step and subtracted from the protein scattering data. All data were processed and analyzed using tools from the ATSAS software package including PRIMUS and CRYSOL.[188–190]

4.2.5 NMR spectroscopy

All spectra were recorded at 25 °C on Bruker 600 MHz, 700 MHz and 800 MHz AVANCE spectrometers, equipped with 5 mm, triple resonance, three-axis gradient probes, or *z*-axis gradient cryoprobes. For three-dimensional NMR experiments, the sample contained 300 μM $^{13}C/^{15}N$-labeled protein in 25 mM NaAc buffer, pH 5.0, 25 mM NaCl, 5 mM DTT, 3 mM NaN_3, and 5% D_2O. For chemical shift assignments, a series of heteronuclear, multidimensional experiments, routinely used in our laboratory, were recorded.[191] Complete ^{1}H, ^{15}N, and ^{13}C backbone resonance

assignments were obtained from 3D HNCACB and HN(CO)CACB spectra, using the program CCPNMR.[192] Weighted chemical shift differences were calculated using the expression: $\Delta\delta=\sqrt{((\Delta\delta_H)^2+(0.15\bullet\Delta\delta_N)^2)}$.

^{15}N R_1 and R_2 relaxation and ^{15}N-{^1H} heteronuclear NOE data were collected on a sample of 100 μM ^{15}N-labeled Mo-WT protein in 25 mM NaAc buffer, pH 5.0, 25 mM NaCl, 5 mM DTT, 3 mM NaN_3, and 5% D_2O, using ^1H-^{15}N HSQC-based pulse sequences at 600 MHz.[193] The R_1 and R_2 experiments employed delays of 0, 100, 150, 200, 300, 500, and 800 ms, and 0, 8, 16, 32, 48, 64, and 80 ms, respectively. Spectra were processed using NMRPipe. R_1 and R_2 relaxation rates were calculated using single exponential fits, and ^{15}N-{^1H} heteronuclear NOE values were calculated using a ratio of experiments recorded with and without ^1H saturation.[194] Rotational correlation times (τ_c) were calculated using the program relax.[195,196] Overlapped resonances and those exhibiting heteronuclear NOE values below 0.7 were omitted from the calculation. 95% confidence intervals of τ_c values were estimated by selecting 1000 subsamples, each with a randomly selected 75% set of rates, calculating τ_c for each.

PRE data were recorded using the single-cysteine variant of Mo-WT, Mo-SAVC (C15S, C25A, C82V). Spectra were recorded at 25 °C using 55 μM ^{15}N-labeled protein in 25 mM NaAc buffer, pH 5.0, 25 mM NaCl, 3 mM NaN_3, and 5% D_2O at 800 MHz. Delays of 0.5, 1.0, 2.0, 4.0, and 8.0 ms were employed. ^{15}N R_2 relaxation rates were calculated using single exponential fits, and ^1H_N-Γ_2 were extracted from the difference between the paramagnetically- and diamagnetically-tagged samples. Results were visualized using the program Visual Molecular Dynamics.[173]

4.2.6 Structure calculation of Mo-WT

Structure calculation of Mo-WT was carried out using XPLOR-NIH version 2.44,[197] subject to restraints from the experimental SAXS and PRE data. Starting models for the calculations were generated by building the MTSL tag onto C167 for each of the 25 solution NMR conformers deposited in the PDB,[22] followed by a round of simulated annealing. During the initial structure generation only XPLOR's molecular geometry terms were applied, and the positions of all atoms in the CVNH and LysM domains were kept fixed while those of the inter-domain linkers (residues 55-61 and 111-117) were unrestrained. The annealing process involved simulating at 10,025 °C for 10 ns before ramping down to 25 °C in 100 °C intervals using a 0.2-ps simulation at each temperature, followed by a 1000-step energy minimization. From each of the 25 original NMR conformers, 25 inter-domain orientations were thereby generated, yielding a total of 625 starting models from which calculations were seeded.

Rotational correlation times used in the back-calculation of $^{1}H_{N}$-Γ_2 rates from molecular coordinates were fixed at the values calculated for the CVNH and LysM domains from the experimental ^{15}N R_1 and R_2 and ^{15}N-{^{1}H} heteronuclear NOE data. SAXS restraints were applied by back-calculating scattering intensity from the molecular coordinates.[198] PRE $^{1}H_{N}$-Γ_2 restraints were derived according to the Solomon–Bloembergen equation and weighted based on the experimental error.[199] These restraints were grouped into intra-domain restraints within the CVNH domain and inter-domain restraints with the LysM domain.

During the production calculations, the CVNH and LysM domain backbone coordinates were fixed, while side chain atoms and all residues in the inter-domain linkers were unrestrained. In each production run, an ensemble of 24 structures was subjected to simulated annealing, with the average back-calculated PRE $^{1}H_{N}$-Γ_2 rates and SAXS intensity restrained to their experimental

values.[198,199] Since residues very close to the MTSL label lack visible amide resonances, $^{1}H_{N}$-Γ_2 rates could not be measured for all residues in the CVNH domain. To derive acceptable conformations of the solvent exposed MTSL tag, which is surrounded by relatively few restraints, simulated annealing was performed in two stages. In the first stage, XPLOR's molecular geometry terms were applied alongside 67 PRE $^{1}H_{N}$-Γ_2 restraints between the MTSL label and backbone amide hydrogens of the CVNH domain. The system was equilibrated at 3000 °C for 100 ps to allow different inter-domain orientations to emerge, compared to those present in the initial structures. Subsequent simulated annealing of the equilibrated system involved cooling down from 3000 °C to 25 °C in 25 °C intervals, with 0.2 ps of simulation at each temperature, followed by a 1000-step energy minimization, after which the coordinates of C167 and the attached MTSL tag were fixed. During the second stage of simulated annealing, XPLOR's molecular geometry terms were applied alongside 39 PRE $^{1}H_{N}$-Γ_2 restraints between the MTSL tag and residues on the LysM domain, as well as a SAXS intensity restraint on the overall system. The system was re-equilibrated at 3000 °C for 100 ps, followed by simulated annealing and energy minimization as described above. Overall, 625 ensembles of 24 structures each were calculated, yielding a total of 15,000 structures. In order to quantify the influence of the inter-domain PRE and SAXS restraints on the resulting structural ensemble, three control calculations were carried out, omitting (i) inter-domain PRE restraints, (ii) SAXS restraints, or (iii) both inter-domain PRE and SAXS restraints from the second stage of simulated annealing.

4.3 RESULTS AND DISCUSSION

The structure and dynamics of flexibly linked multi-domain proteins are particularly challenging to characterize by experimental methods such as X-ray crystallography. Instead, they are well-suited as targets for integrated methods that combine experimental data with computer simulations. Here, we apply such methodology to the two-domain protein MoCVNH3, whose domain-insertion topology sets it apart from linearly connected multi-domain proteins. Our prior structural work by solution NMR and X-ray crystallography[17,22] demonstrated that the two domains have no fixed inter-domain orientation and did not provide details about the nature or distribution of sampled orientations. Here, we used a combination of SAXS, NMR, and MD simulation to characterize the inter-domain orientations of MoCVNH3 as well as the influence of the inter-domain linker lengths on the overall structure and dynamics of the protein. Our results provide extensive data for evaluating the accuracy of simulation models, utilizing this unique system for validating the structure and dynamics of the individual domains in tandem with the overall inter-domain dynamics.

4.3.1 Accessible inter-domain orientations

In our previous work, we investigated the wild-type MoCVNH3 construct, called Mo-WT throughout this manuscript, and several reduced linker-length constructs. Here, we study two of these constructs, Mo-0G and Mo-0v, as well as a new construct, Mo-2G, in detail. In the Mo-0G and Mo-2G constructs, each inter-domain linker is shortened to zero and two glycine residues, respectively (Figure 4.1B). In the Mo-0v construct, three additional residues adjacent to the second

linker are replaced by a single glycine. Mo-0v was successfully crystallized, which proved impossible for the Mo-WT and Mo-0G constructs, despite considerable effort.[17] To eliminate potential confounding factors on the global structure of this two-domain system, we also deleted the four- or six-residue N-terminal cloning artifacts that were present in the proteins studied previously,[17,22] preserving the native amino acid sequence. Consistent with our prior work, a comparison of the ^{1}H-^{15}N HSQC NMR spectra of our new, native, Mo-2G, Mo-0G, and Mo-0v constructs with native Mo-WT revealed only very small chemical shift changes for residues distant from the linkers, demonstrating that the structures of the individual domains are retained in all constructs (Figure 4.1).

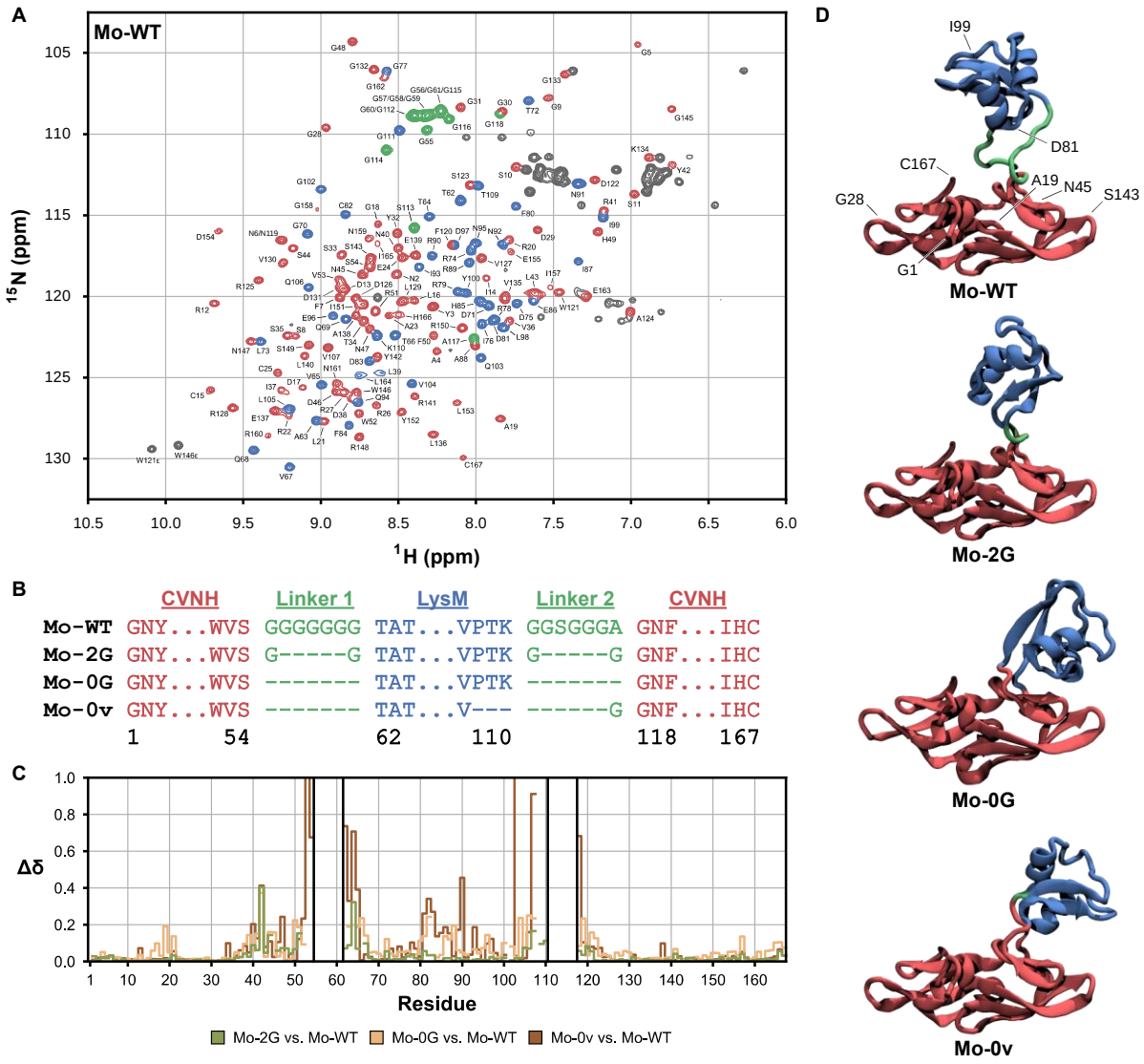

Figure 4.1. MoCVNH3 wild-type (Mo-WT) and reduced linker-length constructs Mo-2G, Mo-0G and Mo-0v. (A) ^{1}H-^{15}N HSQC NMR spectra of Mo-WT. Resonances corresponding to residues in the CVNH domain are shown in red, in the LysM domain in blue, and in the inter-domain linkers in green. Side chain resonances are shown in grey. (B) Amino acid sequences of the MoCVNH3 constructs, highlighting the inter-domain linker regions. Amino acids in the CVNH domain are shown in red, in the LysM domain in blue, and in the inter-domain linkers in green. (C) Chemical shift differences between Mo-WT and Mo-2G, Mo-0G, and Mo-0v constructs. (D) Structures of MoCVNH3 constructs. The Mo-WT structure was previously solved by NMR,[22] and Mo-0v by X-ray crystallography,[17] while Mo-2G and Mo-0G are represented by homology models. Selected residues highlighted in the text are labeled in the Mo-WT structure.

To characterize the influence of the lengths of the inter-domain linkers on the global structure and dynamics of MoCVNH3, we carried out MD simulations of the WT and each of the three reduced linker-length variants (Mo-2G, Mo-0G, and Mo-0v), using two current force field/water model combinations: (i) the AMBER ff15ipq force field (from this point onwards, we will refer to this force field as "ff15ipq") with the SPC/E_b water model,[96,184] and (ii) the CHARMM22* force field with the TIP4P-D water model.[32,102] Both the ff15ipq and CHARMM22* force fields were parameterized to address the issue of overstabilizing salt bridges – a limitation of many other contemporary force fields.[56] The ff15ipq force field is a complete reparametrization, which includes new implicitly polarized atomic charges, new angle parameters, new atomic radii for polar hydrogens, and a greatly expanded set of torsion terms. This force field was developed for use with the SPC/E_b water model, which more accurately reproduces the experimental rotational diffusion of globular proteins, compared to earlier water models.[96] The CHARMM22* force field is a modification of CHARMM22 that includes adjustments to the atomic charges of arginine, aspartate, and glutamate side chains and updates to the backbone torsion parameters. This force field was paired with the TIP4P-D water model, which reduces the oversampling of compact conformations of nonglobular proteins by earlier water models. In total, eight MD simulations were each run for 5 to 10 μs, with an aggregate simulation time of >60 μs (Table 4.1). These simulations provide an excellent opportunity to validate the accuracy with which the two force field/water model combinations reproduce (i) the rotational diffusion of a two-domain protein, expanding on efforts involving single-domain proteins[29,96,184] and (ii) the compactness of a flexibly linked, globular two-domain protein, expanding on efforts involving nonglobular proteins.[102]

Table 4.1. MD simulations of MoCVNH3 constructs

Construct	Force Field	Water Model	Duration
Mo-WT	AMBER ff15ipq	SPC/E_b	10.0 μs
	CHARMM22*	TIP4P-D	7.3 μs
Mo-2G	AMBER ff15ipq	SPC/E_b	10.0 μs
	CHARMM22*	TIP4P-D	5.5 μs
Mo-0G	AMBER ff15ipq	SPC/E_b	10.0 μs
	CHARMM22*	TIP4P-D	5.5 μs
Mo-0v	AMBER ff15ipq	SPC/E_b	10.0 μs
	CHARMM22*	TIP4P-D	5.0 μs

While our earlier solution NMR results lacked detailed information about the inter-domain orientations of the proteins, MD simulations afford the opportunity to efficiently generate large ensembles of orientations. As shown in Figure 4.2, our simulations reveal that ff15ipq/SPC/E_b and CHARMM22*/TIP4P-D yield very different ensembles of inter-domain orientations: while the ff15ipq/SPC/E_b simulations remained in a single inter-domain orientation, the CHARMM22*/TIP4P-D simulations of Mo-WT and Mo-2G sampled a range of inter-domain orientations with more extended conformations. Smaller differences between the two force field/water model combinations are observed for the Mo-0G construct. In contrast, both simulations of Mo-0v yielded nearly identical, more restricted sets of accessible inter-domain orientations.

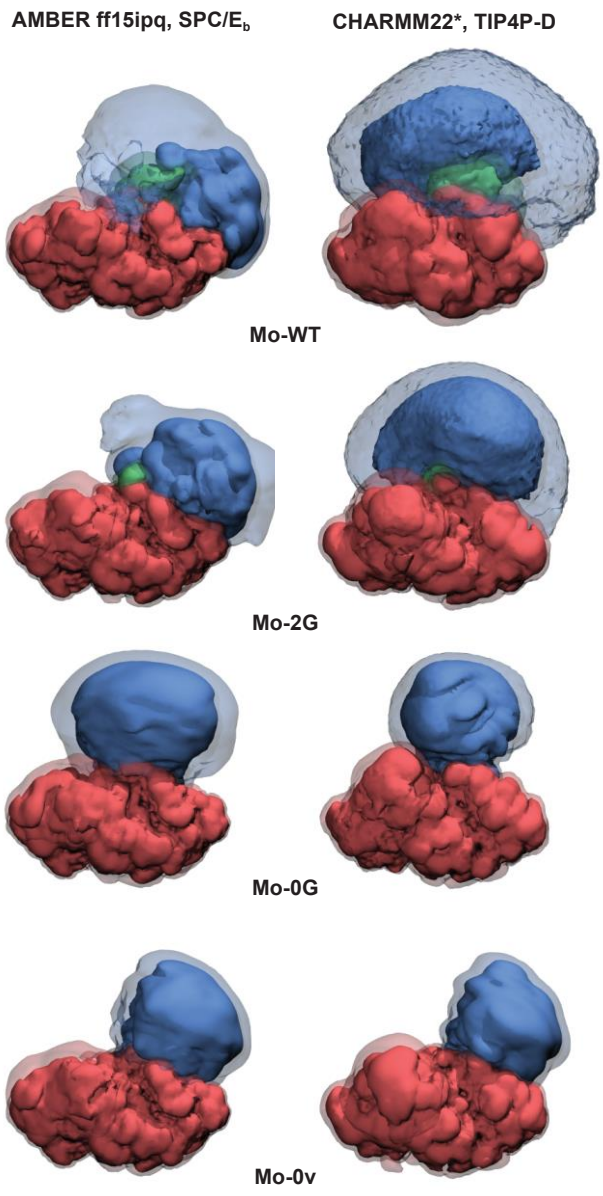

Figure 4.2. Probability distributions of inter-domain orientations sampled in MD simulations for four MoCVNH3 constructs. The CVNH domain is shown in red, the LysM domain in blue, and the inter-domain linkers in green. Trajectories were best fit to the CVNH domain coordinates and the simulation cell was divided into 1-Å^3 bins; solid contours represent bins occupied by a heavy atom for at least 1% of the simulation, while transparent contours represent bins occupied for at least 0.1% of the simulation.

4.3.2 Structural characterization of the CVNH and LysM domains

To assess the ability of the two force field/water model combinations to maintain the integrity of the individual CVNH and LysM domain structures, we monitored the backbone root mean standard deviations (RMSD) from the initial structures over the course of the simulations. All simulations yielded average RMSD values <3.0 Å for both domains. However, inspection of the distributions of RMSD values (Figure 4.3) reveal that significantly larger variations in the RMSD occur over the course of each simulation. In particular, in all of the simulations that were run, and most pronounced for the simulations run with CHARMM22*/TIP4P-D, the RMSD distribution for the CVNH domain is bimodal, sampling two minima with small and large deviations from the starting structure, respectively. In general, the RMSD values increased as the simulations progressed (Figure 4.11), *e.g.*, for Mo-WT and Mo-0G, the RMSD of the CVNH domain remained <3 Å up until 4 μs, after which the deviations increased to ~4 Å. These results underscore the importance of reaching the multi-μs time scale when simulating complex systems such as the MoCVNH3 protein. Although the ff15ipq/SPC/E_b simulations also show a bimodal distribution for the backbone RMSD of the CVNH domain, the RMSD values were consistently <3.0 Å, with the trajectories for Mo-WT, Mo-2G, and Mo-0v settling at lower RMSD values, with few excursions to higher RMSD values.

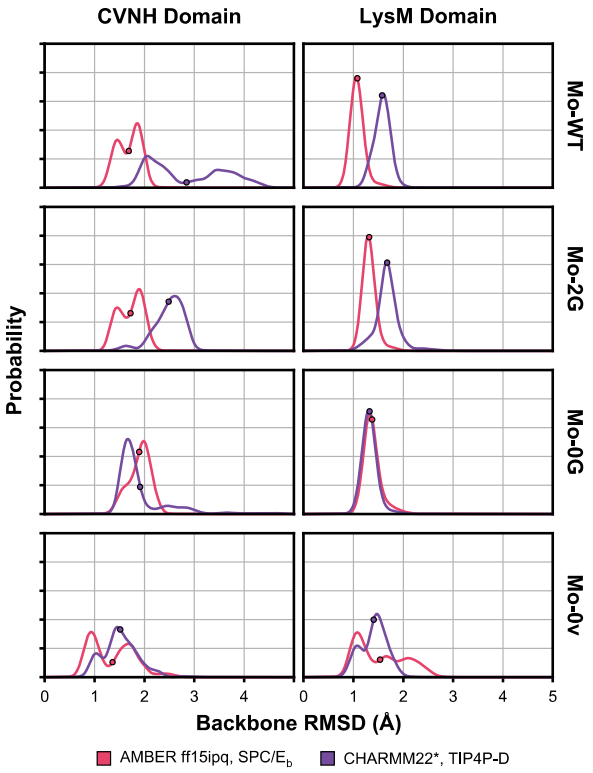

Figure 4.3. Fluctuations in the CVNH and LysM domain coordinates in MoCVNH3 constructs over the course of MD simulations using the AMBER ff15ipq force field/SPC/E_b water model (magenta), and the CHARMM22* force field/TIP4P-D water model (purple), as represented by the distribution of backbone RMSD relative to their initial structures. Average backbone RMSD values are indicated by circles.

For the LysM domain, lower and more tightly distributed RMSD values were observed for the Mo-WT, Mo-2G, and Mo-0G for both field/water model combinations, while RMSD values were more variable for Mo-0v. The overall higher RMSDs observed for the CHARMM22*/TIP4P-D simulations are consistent with observations made during the development of the TIP4P-D water model, which suggested that the implemented increase in the protein-water interaction strength may destabilize the folded states of proteins.[102] Compared to SPC/E_b and most other fixed-charge water models, TIP4P-D increases the relative strength of protein-water interactions vs. protein-protein interactions to reduce the oversampling of compact states of nonglobular proteins that has

resulted from the use of other water models. Our simulations of MoCVNH3 with CHARMM22*/TIP4P-D suggest that this water model may also have the unintentional effect of reducing the sampling of compact (*i.e.,* folded) states of folded proteins, resulting in both a more expanded overall system (Figure 4.2) and less closely packed internal domain structures.

Some curious simulation artifacts were observed for the CVNH domain at the residue level (see Figure 4.12 through Figure 4.28, and Figure 4.1 for the position of highlighted residues within the domain structures). In all eight of our simulations, the N-terminal residues G1 through N6 of the CVNH domain sampled diverse conformations, while from F7 onwards, distributions around a single conformation consistent with the experimental structures dominated along the chain. However, in some of the simulations, a few deviations from the experimental structures were observed as far into the sequence as S10. Also, for the first half of the CVNH domain, the loop spanning residues L16 through A19 sampled multiple conformations in all simulations, while N45 and D46 mostly retained the experimentally determined conformations in simulations run with ff15ipq/SPC/E_b but not with CHARMM22*/TIP4P-D. Within the second half of the CVNH domain, the loop comprising S143 and G144 did not maintain the native conformation in any of our simulations. The greatest differences in CVNH domain coordinates were observed in the simulation of Mo-WT with CHARMM22*/TIP4P-D in which a conformational change occurs in the β-strand that connects the end of the second inter-domain linker to the second half of the CVNH domain, although the antiparallel β-sheet (residues 125 to 151) that makes up most of the second half of CVNH stayed intact (Figure 4.13 and Figure 4.22).

In all eight simulations, the overall structure of the LysM domain was retained more faithfully than that of the CVNH domain. Interestingly, the experimental conformations of loop residues F80 and D81 in the LysM domain were better retained in simulations with

CHARMM22*/TIP4P-D than those with ff15ipq/SPC/E_b. This better retention is most likely due to D81's native left-handed α-helical conformation, which is disfavored by ff15ipq.[184] Interestingly, in seven of the eight simulations, I99 primarily sampled a rare conformation centered at $\Phi \approx 60°$, $\Psi \approx 150°$ while this residue exhibits a PPII conformation in the experimental structures of Mo-WT and Mo-0v. The occurrence of the rare conformation, which is essentially the inverse of the rare "plateau" conformation, may simply be a consequence of the limited functional form of the ff15ipq and CHARMM22* force fields.

To provide insight into why it was possible to crystallize Mo-0v, but not Mo-0G,[17] we compared the inter-domain linker regions of Mo-0G and Mo-0v. In Mo-0G, the first linker between the CVNH and LysM domains, comprising consecutive residues V53, S54, T62, and A63, exhibited diverse conformations in the AMBER ff15ipq/SPC/E_b simulation, while in the CHARMM22*/TIP4P-D simulation, these residues occupied a single conformation. However, in the second inter-domain junction, involving residues P108, T109, K110, G118, and N119, both force fields resulted in diverse conformations for Mo-0G. In contrast, for Mo-0v, in which P108, T109, and K110 have been replaced by a single G117, simulations with both force fields stably retained the conformations that were seen in the crystal structure for the inter-domain junctions. This result suggests that the changes that were introduced into the second linker of Mo-0v resulted in a conformationally more restricted system and may therefore be responsible for its successful crystallization.

4.3.3 Compactness of the two-domain systems

To quantify the influence of the inter-domain linker-lengths on the overall structure of the two-domain MoCVNH3 system, small-angle X-ray scattering (SAXS) curves were measured for Mo-WT and the three reduced linker-length constructs (Figure 4.4). During sample preparation of the Mo-0v protein we noted that this protein was less soluble than the other three constructs, requiring data collection at a lower concentration (0.55 mg/mL compared to 5.0 mg/mL), therefore resulting in noisier data. The lower solubility of Mo-0v may relate to its reduced net charge, as this construct contains one fewer lysine than the larger constructs (Figure 4.1). Earlier work on Mo-0v with the construct, which included a four-residue N-terminal cloning artifact, including a histidine (positively charged at the experimental pH), did not exhibit reduced solubility difference,[17] illustrating that small amino acid changes can greatly influence a protein's behavior. In the present study, we removed the non-native N-terminal amino acids from our protein constructs to eliminate their potential contributions to inter-domain interactions. While the removal of these amino acids had the unfortunate consequence of lowering the solubility of Mo-0v, the data obtained with all the proteins are of sufficient quality for a valid comparison as described below.

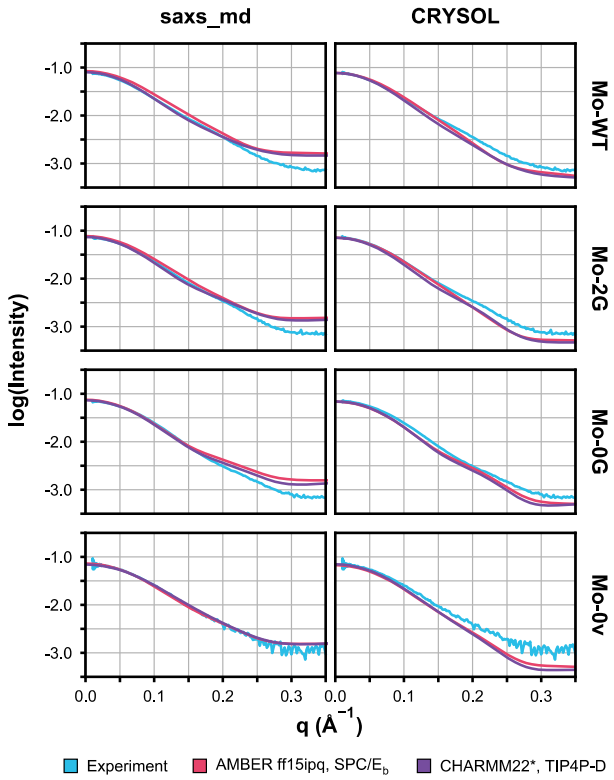

Figure 4.4. Small-angle X-ray scattering intensity of MoCVNH3 constructs measured by experiment (cyan) and back-calculated from MD simulations (magenta, purple). The left panel depicts simulated curves back-calculated using AmberTools' saxs_md program,[187] which includes explicit water molecules for calculating the scattering, and the right panel shows simulated curves back-calculated with ATSAS' CRYSOL program,[188] which represents solvent implicitly. The scale of the y-axis is arbitrary; shaded regions represent 95% confidence intervals for the experimental and simulated values; the larger uncertainty in the experimental scattering for the Mo-0v construct arises from the need to collect the data at lower concentration due to the protein's lower solubility.

The experimental SAXS data offer the opportunity to validate our MD simulations' modeling of the overall structural of the two-domain system. To this end, we back-calculated SAXS curves from the simulation coordinates, using two different methods: (i) AmberTools' saxs_md program,[187] which explicitly includes the coordinates of surrounding water molecules in the calculation, and (ii) ATSAS' CRYSOL program,[188] which implicitly accounts for the scattering

of surrounding water molecules. As shown in Figure 4.4, for q values below 0.2 $Å^{-1}$, the two methods yield results that are broadly similar to one another and to experiment. Beyond 0.2 $Å^{-1}$, the saxs_md program consistently yields higher scattering intensity than observed experimentally, while CRYSOL yields lower scattering intensity. Since we were interested in capturing differences between the four constructs, we calculated the differences in the corresponding scattering intensities (Figure 4.29). The small, systematic difference between the two calculation methods is of little consequence, since for q >0.2 $Å^{-1}$ the curves for the MoCVNH3 constructs are indistinguishable. For q <0.2 $Å^{-1}$, the two methods of back-calculation yield similar results, suggesting that, for the region of q measured here, the computationally more expensive explicit-solvent saxs_md calculation does not provide a tangible benefit over the less demanding CRYSOL calculation. Qualitatively, the CHARMM22*/TIP4P-D simulations reproduce the experimental trends more accurately than ff15ipq/SPC/E_b. Overall, the above results suggest that the global structure of the MoCVNH3 two-domain system is more accurately represented by using CHARMM22*/TIP4P-D than ff15ipq/SPC/E_b.

A key structural parameter that can be calculated from the measured SAXS curves is the radius of gyration (R_g), which reflects the compactness of the protein. Since all the MoCVNH3 constructs are similar in overall mass, R_g provides a means by which the effect of the different linker lengths on the population of extended vs. collapsed conformations in the two-domain system can be assessed. As shown in Figure 4.5, the R_g for Mo-WT, Mo-2G, and Mo-0G, calculated from the experimental SAXS curves, exhibit a clear and intuitively expected decrease of R_g, with reduced inter-domain linker length. While the data for Mo-0v are too noisy to confidently differentiate its R_g from that of Mo-0G and Mo-2G, the R_g of Mo-0v is statistically distinguishable

from that of Mo-WT, and the center of the confidence interval lies just below that of Mo-0G, consistent with its slightly shorter second inter-domain linker.

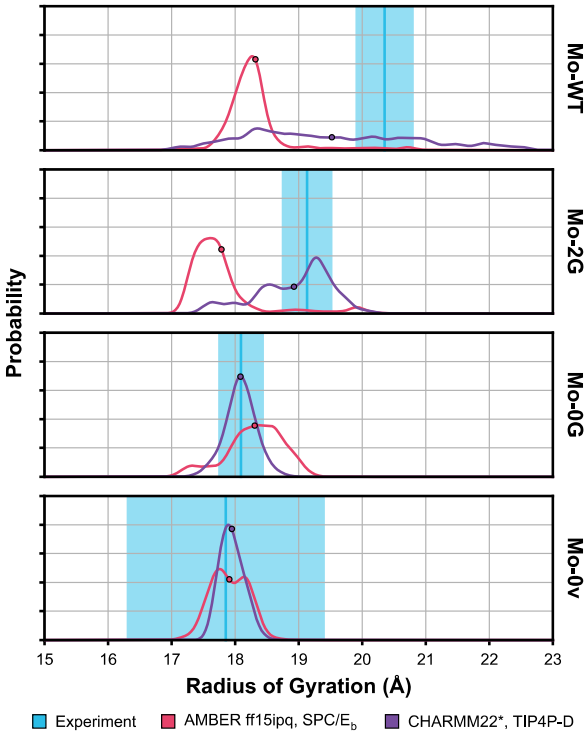

Figure 4.5. Radius of gyration (R_g) of MoCVNH3 constructs calculated from experimental SAXS intensity (cyan) and from conformations sampled in MD simulations (magenta, purple). Cyan shaded regions represent 95% confidence intervals on the experimental values. Average R_g values are indicated by circles.

R_g can be calculated straightforwardly from the MD simulation coordinates, providing further validation of the accuracy of the constructs' global simulated ensembles. The distributions of R_g, sampled over the course of each simulation (Figure 4.5), are consistent with the observations about the back-calculated SAXS curves: the CHARMM22*/TIP4P-D simulations more accurately reproduce the experimental R_g value, compared to the ff15ipq/SPC/E_b simulations. Both the ff15ipq/SPC/E_b and CHARMM22*/TIP4P-D trajectories yielded R_g values of ~18 Å for Mo-0v and Mo-0G, in excellent agreement with experiment. The simulations of Mo-2G and Mo-WT, run

with ff15ipq/SPC/E_b, yielded R_g values similar to those of Mo-0G and Mo-0v, while those run with CHARMM22*/TIP4P-D resulted in larger R_g values, in much better agreement with experiment. However, the R_g value of Mo-WT obtained with CHARMM22*/TIP4P-D is still ~1 Å below the experimental value and fluctuated significantly over the course of the simulation (Figure 4.30). The observed broad distribution suggests that longer simulations than performed here (7.3 μs) may be necessary to achieve convergence for Mo-WT. Both experiment and simulation yielded R_g values of ~18 Å for Mo-0G and Mo-0v; our simulations of Mo-WT sampled conformations this compact when the domains were in contact, as well as conformations with R_g of up to 23 Å when the domains were not in contact. The experimental data for Mo-WT shows an R_g value of 20.4 Å, which is 12% larger than those of Mo-0G and Mo-0v as a result of more frequent sampling of extended conformations.

4.3.4 Dynamical properties of wild-type MoCVNH3 (Mo-WT)

The global structural information obtained from SAXS can be complemented with single-residue and single-domain dynamics information, accessible by NMR relaxation approaches. In particular, the ratio between the ^{15}N transverse (R_2) and longitudinal (R_1) relaxation of the backbone amide resonances provides a measure of the system's rotational diffusion in solution, the isotropic rotational correlation time, τ_c. NMR relaxation data were collected for the Mo-WT construct, and the two domains exhibited characteristic R_2/R_1 ratios of ~9 and ~5 for the CVNH and LysM domains, respectively (Figure 4.6), corresponding to τ_c values of 8.7 and 6.7 ns. This difference in correlation time confirms that both domains tumble essentially independently in solution.

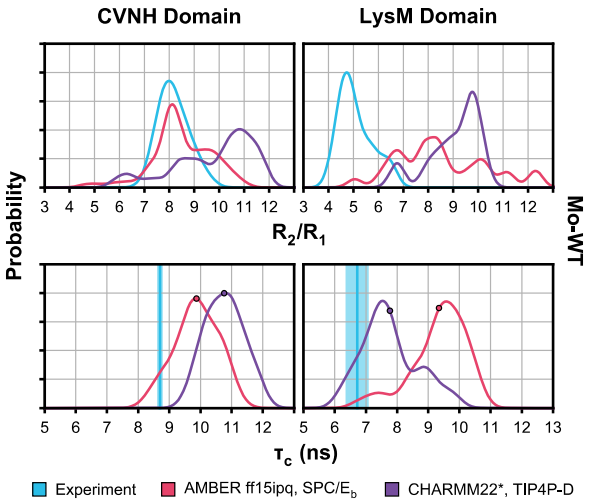

Figure 4.6. Rotational diffusion of CVNH and LysM domains of Mo-WT calculated from experimental NMR relaxation (cyan) and MD simulation (red, purple). Top, distribution of R_2/R_1 relaxation ratios for residues in the CVNH (left) and LysM domain (right). Bottom, distribution of calculated rotational correlation times (τ_c) for the CVNH (left) and LysM domains (right). Cyan shaded regions represent 95% confidence intervals of the experimental values. Average values are indicated by circles.

To validate the accuracy of the two MD simulations for Mo-WT, ^{15}N R_1 and R_2 relaxation rates were back-calculated from the motions of the backbone amide N-H vectors over the course of the simulations. We previously noted that the ff15ipq/SPC/E_b force field/water model combination yielded accurate rotational diffusion times for single-domain globular proteins, while the CHARMM22*/TIP4P-D combination yielded less accurate results.[184] In contrast to the well-defined distributions of the experimental R_2/R_1 ratios, the R_2/R_1 ratios back-calculated from the simulations exhibited much greater variability for individual residues within each domain (Figure 4.31). Inspection of the rotational correlation times extracted from the simulation of Mo-WT with ff15ipq/SPC/E_b resulted in similar τ_c values of 9-10 ns for both domains. This implies that in the ff15ipq/SPC/E_b simulations the dynamics of the two domains are too tightly coupled and is consistent with our observation that the ff15ipq/SPC/E_b conformational ensemble is too compact

(Figure 4.5), adopting only a single inter-domain orientation (Figure 4.1). The CHARMM22*/TIP4P-D simulation yielded τ_c values of ~11 ns and 7 ns, capturing the difference in rotational diffusion between the domains. However, the τ_c value of the CVNH domain is somewhat higher than the experimentally measured one, consistent with our prior results for single-domain proteins using this force field/water model combination.[184]

Tracking τ_c over the course of the simulation reveals that for the first microsecond of the ff15ipq/SPC/E_b simulation the τ_c of the LysM domain was relatively accurate, but became worse at ~700 ns when the two domains collapsed onto each other and remained in a single inter-domain orientation for the remainder of the simulation (Figure 4.32). It therefore appears that the ff15ipq/SPC/E_b combination provided accurate rotational diffusion for the conformations that were sampled, but that the distribution of sampled conformations was inaccurate.

4.3.5 Preferred inter-domain orientations of Mo-WT

To determine the distribution of inter-domain orientations within the Mo-WT protein, paramagnetic relaxation enhancement (PRE) measurements were carried out with a single-cysteine variant of Mo-WT. PREs represent the increased relaxation of nearby nuclear spins around a paramagnetic moiety, resulting in broadening of the associated resonances. The large magnetic moment of the unpaired electron causes a large effect, and PREs can be observed over distances up to 35 Å. The PRE effect scales as $1/r^6$, with r the distance between the unpaired electron in the paramagnetic center and the affected nucleus. Further, in systems that exchange rapidly between different conformations, the measured PREs are the population-weighted averages of the PREs for all sampled conformations, allowing even transient, low-population contacts to be captured.[200] Finally, PREs can be back-calculated from known structures, permitting integration into structure

calculations and validation of candidate models,[199] rendering the PRE approach the ideal methodology for the characterization of distributions.

To probe whether contacts between the two domains of Mo-WT can be captured by PREs, we attached the paramagnetic MTSL tag to a cysteine residue in the CVNH domain and measured PREs on residues in the LysM domain. Since Mo-WT contains four cysteines, none of which are involved in disulphide bonds in the native structure,[22] it was necessary to remove all but one cysteine for single site spin-labeling. This was achieved by introducing C15S and C25A mutations into the CVNH domain and the C82V mutation into the LysM domain, leaving C167 at the C-terminus exclusively available for attachment of the paramagnetic tag (Figure 4.9 and Figure 4.33). To verify that attachment of the MTSL tag did not affect the structure of the protein, we recorded the ^{1}H-^{15}N HSQC spectrum of Mo-SAVC tagged with the diamagnetic analog of MTSL at C167 (Figure 4.7A). Compared to the spectrum of untagged Mo-SAVC, only very small chemical shift changes were noted for amino acids close to the C167 attachment site, but not elsewhere (Figure 4.7B, D). Interestingly, several of the affected resonances exhibited doubling, suggesting that attachment of the tag results in two slightly different conformations its vicinity. In the paramagnetically-tagged species, however, the equivalent resonances were broadened beyond detection and were therefore not included in the analysis.

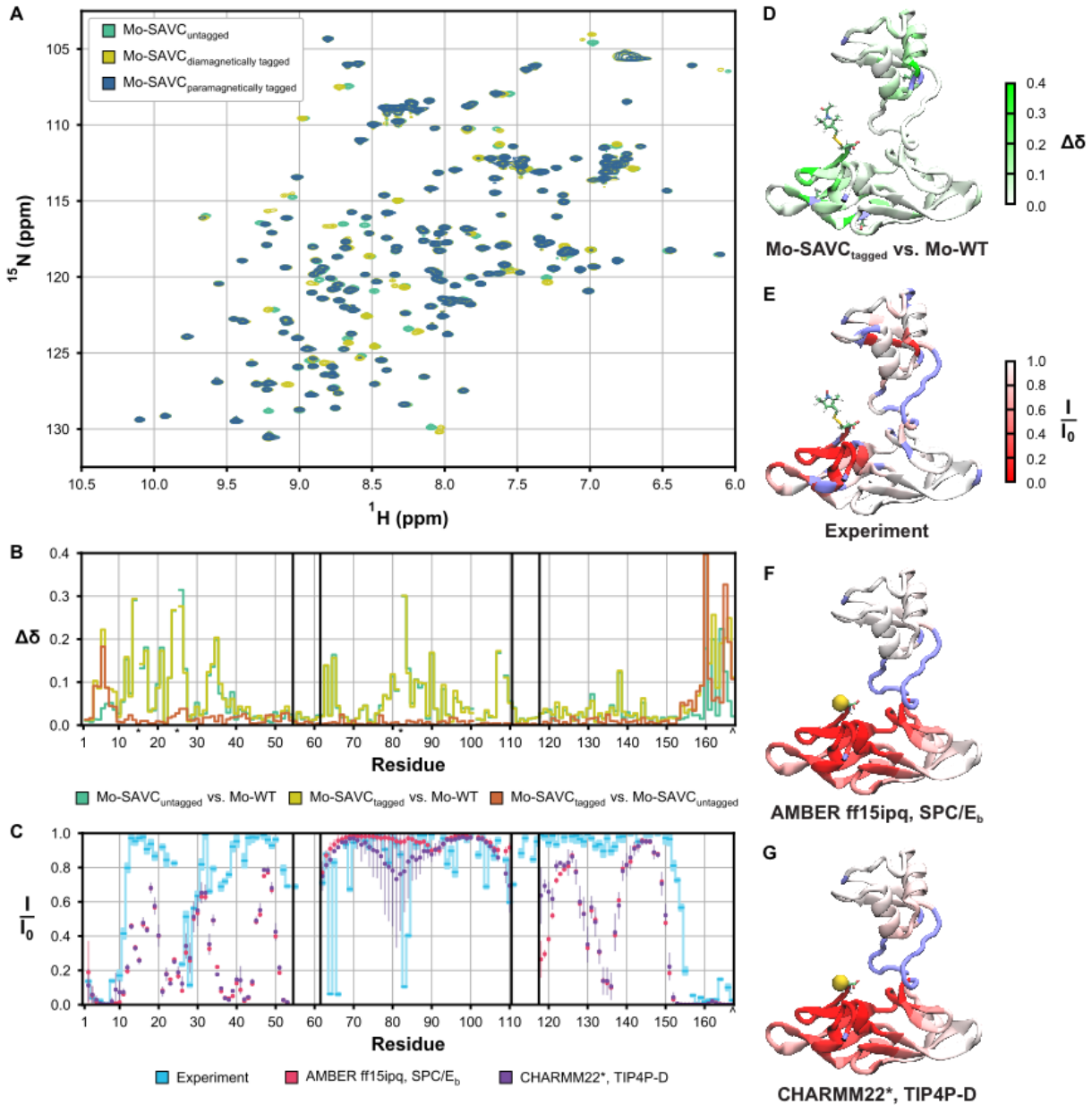

Figure 4.7. PRE data for Mo-WT. (A) Superposition of the ^{1}H-^{15}N HSQC NMR spectra of untagged Mo-SAVC (green) and Mo-SAVC with paramagnetic (teal) or diamagnetic (gold) tags attached to C167. (B) Chemical shift differences between Mo-WT and Mo-SAVC, and between tagged and untagged Mo-SAVC. The positions of the changed cysteines C15S, C25A, and C82S are marked with asterisks, and C167 to which the tags are attached is marked with a caret. (C) Ratio of peak intensities for the paramagnetically- and diamagnetically-tagged proteins, measured experimentally (cyan) and calculated from MD simulations (red, purple). Cyan shaded region represents 95% confidence interval of the experimental ratios. Since the MTSL tag was not present in the MD simulations,

simulated intensities were calculated based on the distances between each backbone amide hydrogen and Sγ of C167. (D) Chemical shift differences between Mo-WT and diamagnetically-tagged Mo-SAVC, mapped onto the structure of MoCVNH3, using a white to green gradient to represent the degree of difference. Residues for which chemical shift differences are not available are shown in blue. The substituted amino acids C15, C25, and C82, as well as C167 with the tag are shown in licorice representation. (E) Experimental and back-calculated (F) AMBER ff15ipq/SPC/E_b and (G) CHARMM22*/TIP4P-D ratios of peak intensities mapped onto the experimental NMR structure, using a red to white gradient to represent the intensity ratio. Residues for which intensity ratios are not available are shown in blue. C167 with the tag is shown in licorice representation.

PREs were quantified by comparing resonance intensities in the spectra of paramagnetically- and diamagnetically-tagged samples (Figure 4.7C). For the CVNH domain, the experimentally determined intensity ratios exhibited a clear dependence on the distance between each residue and the MTSL tag (Figure 4.7E). For residues in the LysM domain, the amide resonances of T64, T66, D83, and F84 show strikingly lower intensity ratios than resonances of other residues in this domain. Given that T64 and T66 lie on one side of the LysM domain and D83 and F84 lie on the other side of the domain, it is impossible for all four of these residues to be simultaneously close to the MTSL tag. Thus, the Mo-WT protein must be exchanging between two different orientations with different sides of the LysM domain transiently approaching the MTSL tag.

To determine whether our MD simulations of Mo-WT captured these two different orientations, PRE intensity ratios were back-calculated from the sampled conformations. Since the MTSL tag was not included in the simulation model, the Sγ atom of C167 was used as a proxy. Within the CVNH domain, agreement between the simulated and experimental intensity ratios was observed only for residues far away from C167 (Figure 4.7C, F, G). This indicates that the approximation of the paramagnetic group's location in our simulations is insufficiently accurate

for residues close to the site, where small inaccuracies in distance and orientation have large effects on the back-calculated PREs. However, the imprecision of the paramagnetic tag's location has a smaller effect on ratios calculated for residues in the LysM domain, which are on average further away from C167. For the first contact site on the LysM domain, containing T64 and T66, neither simulation reproduces the experimental ratios, and none of these two amino acids gets close to C167. However, for the second contact site, including D83 and F84, the simulation with CHARMM22*/TIP4P-D results in smaller intensity ratios for these and several nearby residues centered around D81. Although it is possible that the simulated conformations in which D81 approaches C167 are representative of the experimental conformations that are responsible for the low intensity ratios of D83 and F84, the large standard errors of the back-calculated inter-domain PREs suggest that even longer simulations may be required to obtain reliable distributions of inter-domain orientations.

To quantitatively link the two contact sites on the LysM domain to the global structure of Mo-WT, we measured $^{1}H_N$-Γ_2 PRE rates, representing the R_2 relaxation induced by the paramagnetic tag, and incorporated them alongside our SAXS data as restraints in the calculation of a structural ensemble using XPLOR-NIH.[197,201] The calculation was seeded from the prior NMR structure,[18] in which the individual domain structures of CVNH and LysM were retained. A total of 15,000 structures were calculated, and the results yielded good agreement with the experimental restraints. The Q-factors for the PRE $^{1}H_N$-Γ_2 rates are 0.41 for CVNH domain residues and 0.52 for LysM domain residues, while the back-calculated SAXS curves resulted in a X^2 value of 0.04. The average R_g of the calculated structural ensemble is 20.0 Å, which is within the 95% confidence interval of the value calculated from the experimental SAXS data using Guinier analysis (20.4 Å).

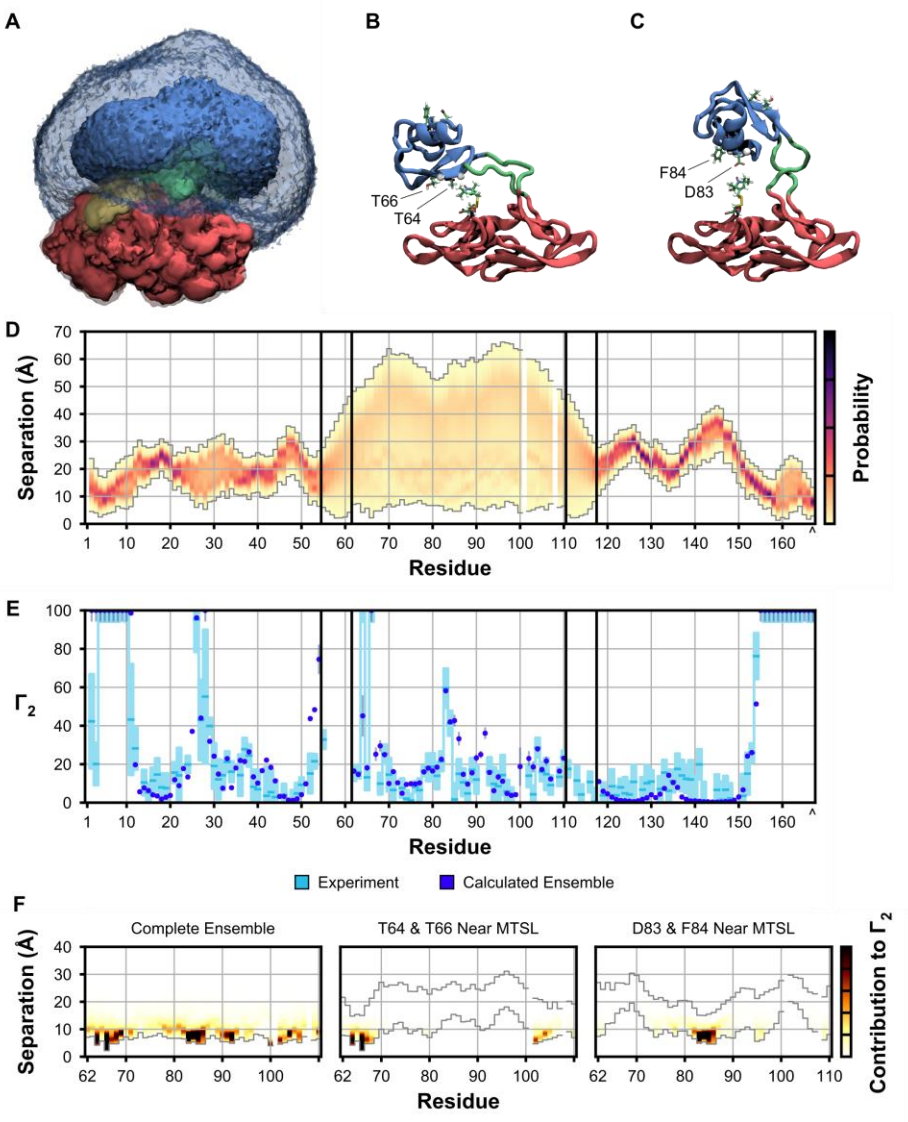

Figure 4.8. Structural ensemble of Mo-WT based on experimental SAXS and PRE data. (A) Probability distributions of inter-domain orientations in the calculated structural ensemble. The CVNH domain is shown in red, the LysM domain in blue, the inter-domain linkers in green, and MTSL paramagnetic tag in yellow. Structures were best fit to the CVNH domain coordinates, and the simulation cell was divided into 1-Å^3 bins; solid contours represent bins occupied by a heavy atom for at least 1% of the ensemble, while transparent contours represent bins occupied for at least 0.1% of the ensemble. (B) A representative structure illustrating the approach between the MTSL tag and T64 and T66. (C) A representative structure illustrating the approach between the MTSL tag and D83 and F94. (D) Distributions of inter-atomic separation between the backbone amide hydrogen of each residue and the nitroxide radical of MTSL. (E) Γ_2 relaxation measured experimentally (cyan) and back-calculated from the structural ensemble

(blue). Cyan shaded region represents 95% confidence interval of the experimental ratios. For clarity, residues whose Γ_2 exceeded 100 are shown as 100. (F) Contribution of mutually exclusive subsets of conformations to back-calculated Γ_2. For both the first (T64 and T66) and second contact sites (D83 and F84), nearly all contribution to the back-calculated Γ_2 comes from a subset of structures comprising ~5 % of the total ensemble.

Within the calculated, experimental data-derived ensemble, the probability distribution of inter-domain orientations shows that a wide range of orientations is sampled by the system (Figure 4.8A). Comparison of this experiment-driven ensemble with the two ensembles derived from MD simulations alone (Figure 4.2) revealed that CHARMM22*/TIP4P-D reproduced the sampled inter-domain orientations in the experiment-driven ensemble more accurately than ff15ipq/SPC/E_b. This demonstrates that the rebalancing of protein/protein and protein/water nonbonded dispersion interactions in the TIP4P-D water model has benefits for systems beyond disordered peptides and proteins, such as the one investigated here.

Table 4.2. Structure calculation of Mo-WT

Restraints Applied	CVNH Domain PRE		LysM Domain PRE		SAXS
	Q-Factor	R^2	Q-Factor	R^2	X^2
Inter-Domain PRE and SAXS	0.41	0.94	0.56	0.67	0.04
Inter-Domain PRE only	0.41	0.94	0.52	0.70	0.95
SAXS only	0.41	0.94	0.67	0.53	0.04
None	0.41	0.94	0.76	0.47	0.13

The distributions of sampled inter-atomic distances between each backbone amide hydrogen and the nitroxide group of MTSL (Figure 4.8D) reveal that in most structures, amino acids in the LysM domain are relatively far away from MTSL. However, for distances <25 Å, it is possible to discern subsets of structures for which parts of the LysM domain are closer to the MTSL tag. Indeed, the back-calculated PRE ^1H_N-Γ_2 rates (Figure 4.8E) reveal that the two contact sites on the LysM domain, identified above from the paramagnetic/diamagnetic intensity ratios (Figure 4.7C, E), are captured in the ensemble. The high Γ_2 of residues D83 and F84 are accurately

reproduced, and residues T64 and T66 also yield high Γ_2. This finding is particularly gratifying, since no restraints on the Γ_2 of T64 and T66 were applied during the calculation. Although their low intensity ratio indicated that their Γ_2 rate had to be high, the peak intensities were too low to confidently measure a Γ_2 rate, which could be converted into a restraint. However, the Γ_2 restraints of surrounding residues clearly were sufficient to capture this contact site in the calculated ensemble. The subset of structures in which the backbone amide hydrogens of both T64 and T66 are within 13 Å of MTSL's nitroxide group comprises 4.0% of all structures and is responsible for over 80% of the ensemble's calculated Γ_2 for these two residues (Figure 4.8B, F). The analogous subset for D83 and D84 (Figure 4.8C, F) comprises 6.6% of structures, which are responsible for over 90% of the calculated Γ_2 of these residues. The two subsets are mutually exclusive, and each one does not contribute to the Γ_2 values of the other set's contact site. This result vividly demonstrates and supports previous findings about the significant influence of low-population states on measured PRE Γ_2 rates.[200]

In order to dissect the influence of the individual inter-domain PRE and SAXS restraint terms on the ensemble's distribution of inter-domain orientations, we calculated three control ensembles, in which the inter-domain orientations were restrained by (i) only inter-domain PRE restraints, (ii) only the SAXS restraint, or (iii) neither. The Q-factors and X^2 values for all three ensembles are summarized in Table 4.2 (the distributions are depicted in Figure 4.34). Unexpectedly, the ensemble that included neither the inter-domain PRE nor SAXS restraints yielded a surprisingly low SAXS X^2 value of 0.13, suggesting that a repulsion term which ensures that the domains cannot overlap spatially and geometric terms that account for the linker lengths reproduce the conformational space accessible to the domains relatively well. However, this unrestrained ensemble does not satisfactorily reproduce the LysM domain Γ_2 values, resulting in

a Q-factor of 0.76. It also does not capture the two contact sites on the LysM domain. This illustrates that, although no fixed inter-domain orientation is present for the two domains, measurable differences in population between the accessible orientations for the ensembles can be discerned.

4.4 CONCLUSIONS

In this work, we have characterized the global structure and dynamics of the flexibly linked domain-insertion protein MoCVNH3, using experimental NMR and SAXS studies in combination with μs time scale MD simulations. To evaluate the influence of inter-domain linker length on the properties of the system, we studied a series of reduced linker-length constructs, in which the two domains ultimately become locked into a single inter-domain orientation. For four tested linker lengths, the global structural properties of the systems were measured by SAXS, and results were compared to MD simulations, testing the ff15ipq/SPC/E_b and CHARMM22*/TIP4P-D force field/water model combinations. We found that while ff15ipq/SPC/E_b more accurately retained the experimental structures of the individual domains, only CHARMM22*/TIP4P-D reproduced the observed increase in the radii of gyration (R_g) for increasing linker length. The inter-domain orientations of the wild-type protein were evaluated by PRE measurements, which identified two mutually exclusive contact sites between the CVNH and LysM domains. Finally, we used our SAXS and PRE data in combination to calculate an overall structural ensemble. Our results show that while no fixed inter-domain orientation exists for the two domains of MoCVNH3, measurable differences in population for the accessible orientations can be discerned.

The above findings are valuable in the context of integrative structural biology, where through combinations of various experimental data and computer models one aspires to derive a more comprehensive view of structure and dynamics than is accessible from either experiment or computation alone. Naturally, all computational models are subject to the accuracy of the selected molecular mechanical force fields, which, although quite robust, still possess considerable room for improvement. Recent advancements in model development, such as implemented in the Implicitly Polarized Q (IPolQ) and ForceBalance approaches, reduce the necessary time consuming efforts in each round of improvement.[92,184] Coupled to advances in computer hardware and algorithms, which increasingly enable longer simulations of larger systems,[25,78,79,202,203] ever more complex systems will become accessible to simulation. We suggest that the joint simulation/experimental study of MoCVNH3 reported here provides a valuable benchmark towards this end, in particular for the characterization of structural and dynamical properties of multi-domain proteins.

Overall, the characterization of the structure and dynamics of the two-domain MoCVNH3 protein is, to our knowledge, the most in-depth biophysical characterization of a domain-insertion protein system and illustrates the value of integrating a synergistic combination of NMR, SAXS, and long time scale atomistic simulations for characterizing structural ensembles of flexibly linked multi-domain systems.

4.5 ACKOWLEDGEMENTS

This work was supported by National Institutes of Health grants RO1GM115805 (to L.T.C.) and RO1GM080642 (to A.M.G.). K.T.D. was supported by a National Institutes of Health training

grant GM088119 and a University of Pittsburgh Andrew Mellow Fellowship. Anton computer time was provided by the National Resource for Biomedical Supercomputing (NRBSC), the Pittsburgh Supercomputing Center (PSC), and the BTRC for Multiscale Modeling of Biological Systems (MMBioS) through Grant P41GM103712-S1 from the National Institutes of Health. The Anton computer at NRBSC/PSC was generously made available by D. E. Shaw Research. Local computational resources, including GPU nodes, were provided by a National Science Foundation MRI Award CNS-1229064 and the University of Pittsburgh's Center for Research Computing. The SAXS data were collected through the SAXS Core Facility of the Center for Cancer Research, National Cancer Institute. Time on the shared scattering beamline 12-ID-B was allocated under the PUP-24152 agreement between the National Cancer Institute and Argonne National Laboratory (Lemont, IL, USA). The Advanced Photon Source, a US Department of Energy Office of Science User Facility, is operated by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. We thank Mike Delk for NMR technical support, Drs. Rieko Ishima and Zhaoyong Xi for help setting up NMR relaxation experiments, and David Case and Charles Schwieters for enlightening discussions.

4.6 SUPPORTING INFORMATION

4.6.1 Supporting figures

Figure 4.9. Amino acid sequences of MoCVNH3 constructs. Residues in the CVNH domain are shown in red, in the LysM in blue, respectively, and in the inter-domain linkers in green. In the Mo-SAVC sequence, the C15S, C25A, and C82V amino acid changes as well as C167, to which the paramagnetic or diamagnetic tags were attached, are shown in bold type.

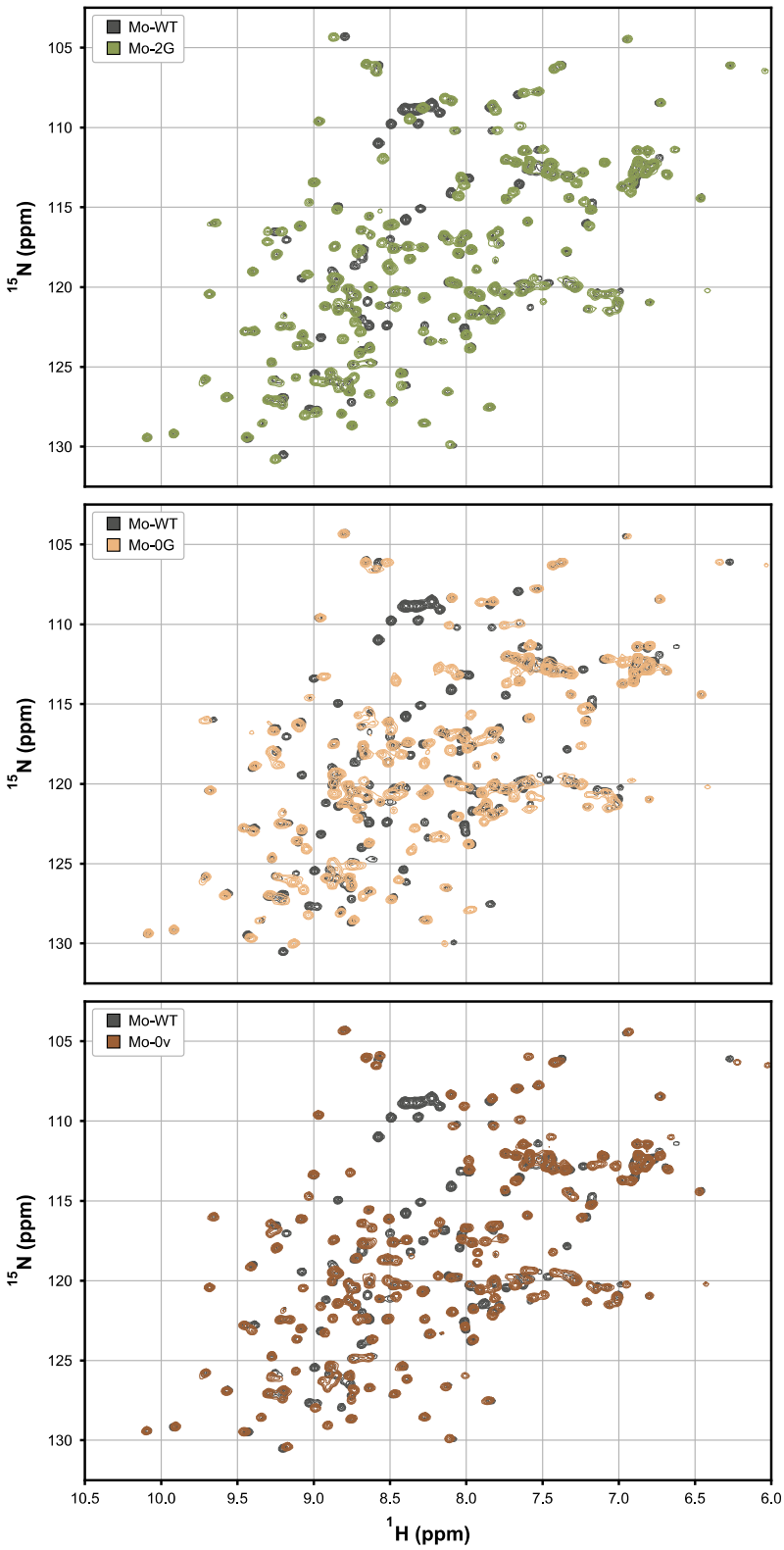

Figure 4.10. Superposition of the ^{1}H-^{15}N HSQC NMR spectra of Mo-2G (olive), Mo-0G (peach), Mo-0v (brown) and Mo-WT (black).

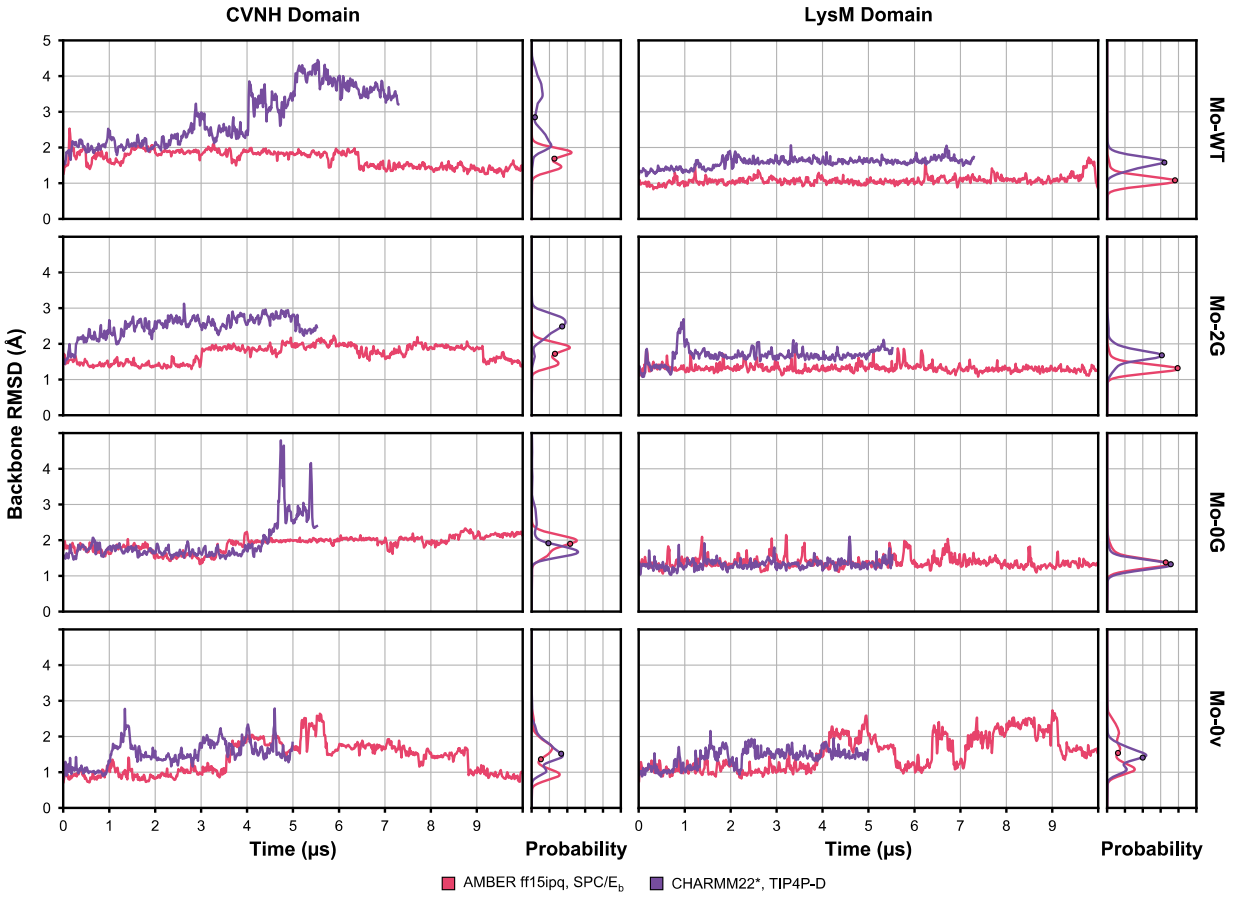

Figure 4.11. Backbone RMSD of CVNH and LysM domain atoms of MoCVNH3 constructs over the course of simulations with ff15ipq/SPC/E_b (magenta) and CHARMM22*/TIP4P-D (purple). The probability distributions of sampled values are shown in the right panels, with the average values indicated by circles.

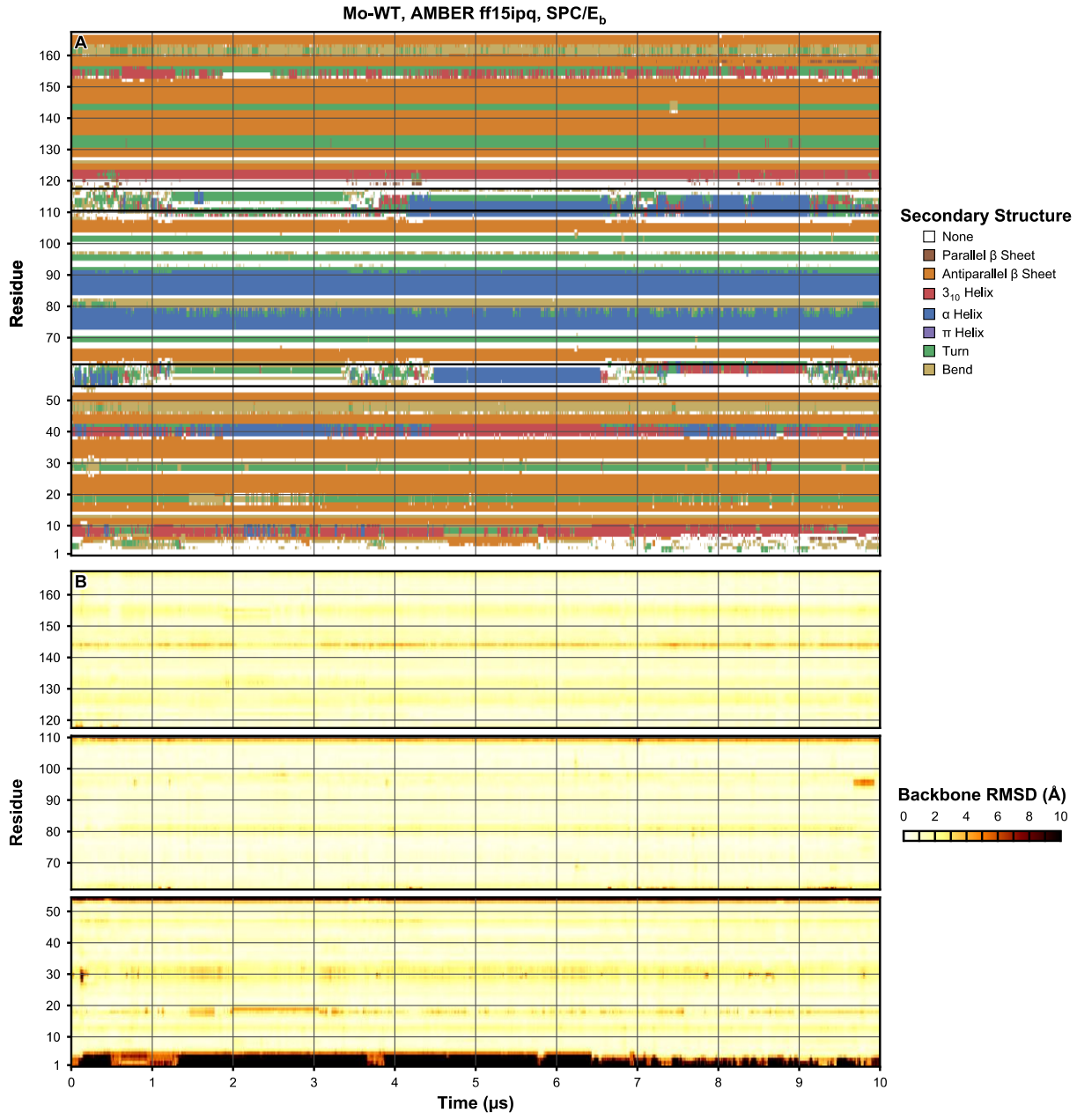

Figure 4.12. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the NMR structure (PDB code 2L9Y) (B) of Mo-WT over the course of a 10-μs simulation with ff15ipq/SPC/E_b.

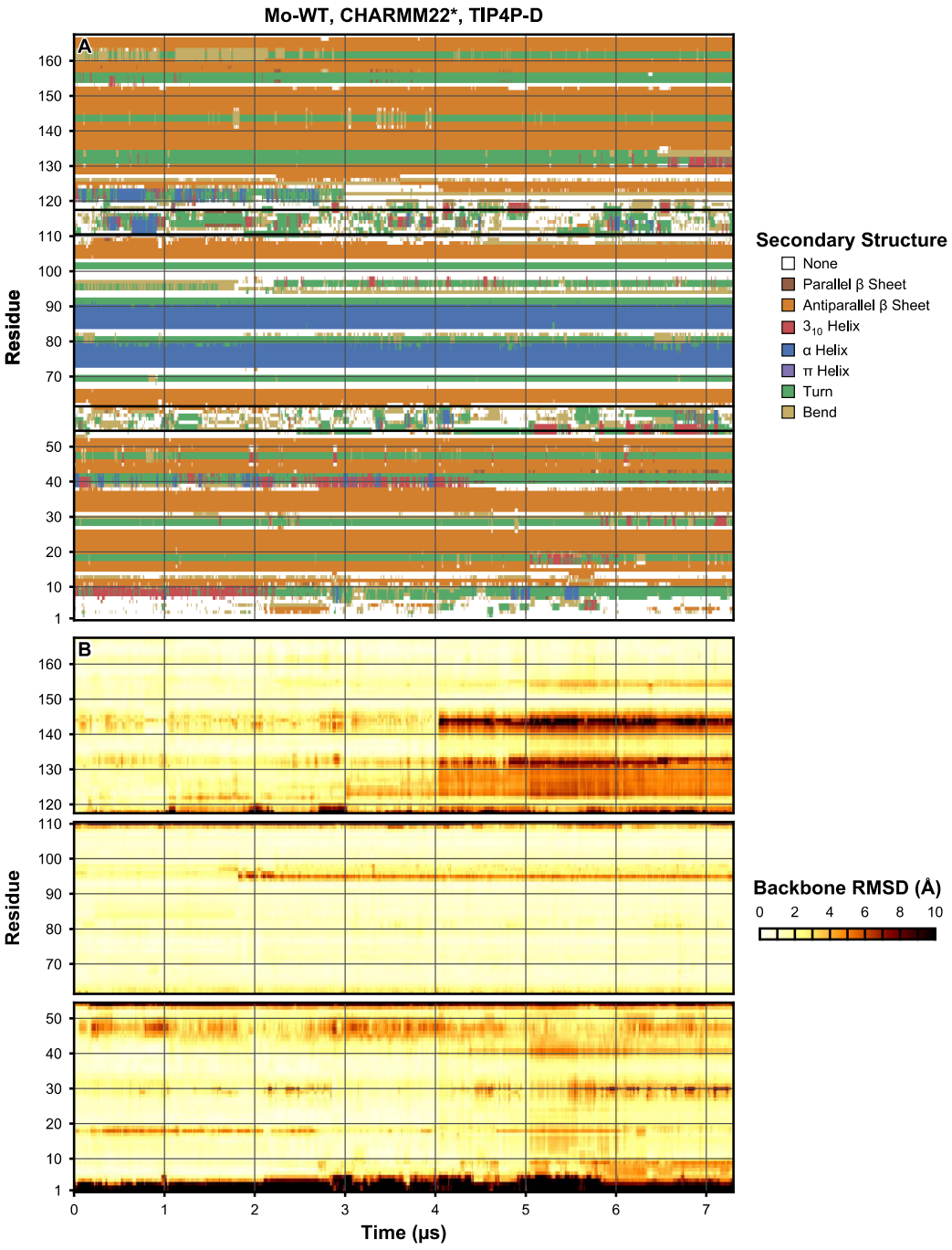

Figure 4.13. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the NMR structure (PDB code 2L9Y) (B) of Mo-WT over the course of a 7.3-μs simulation with CHARMM22*/TIP4P-D.

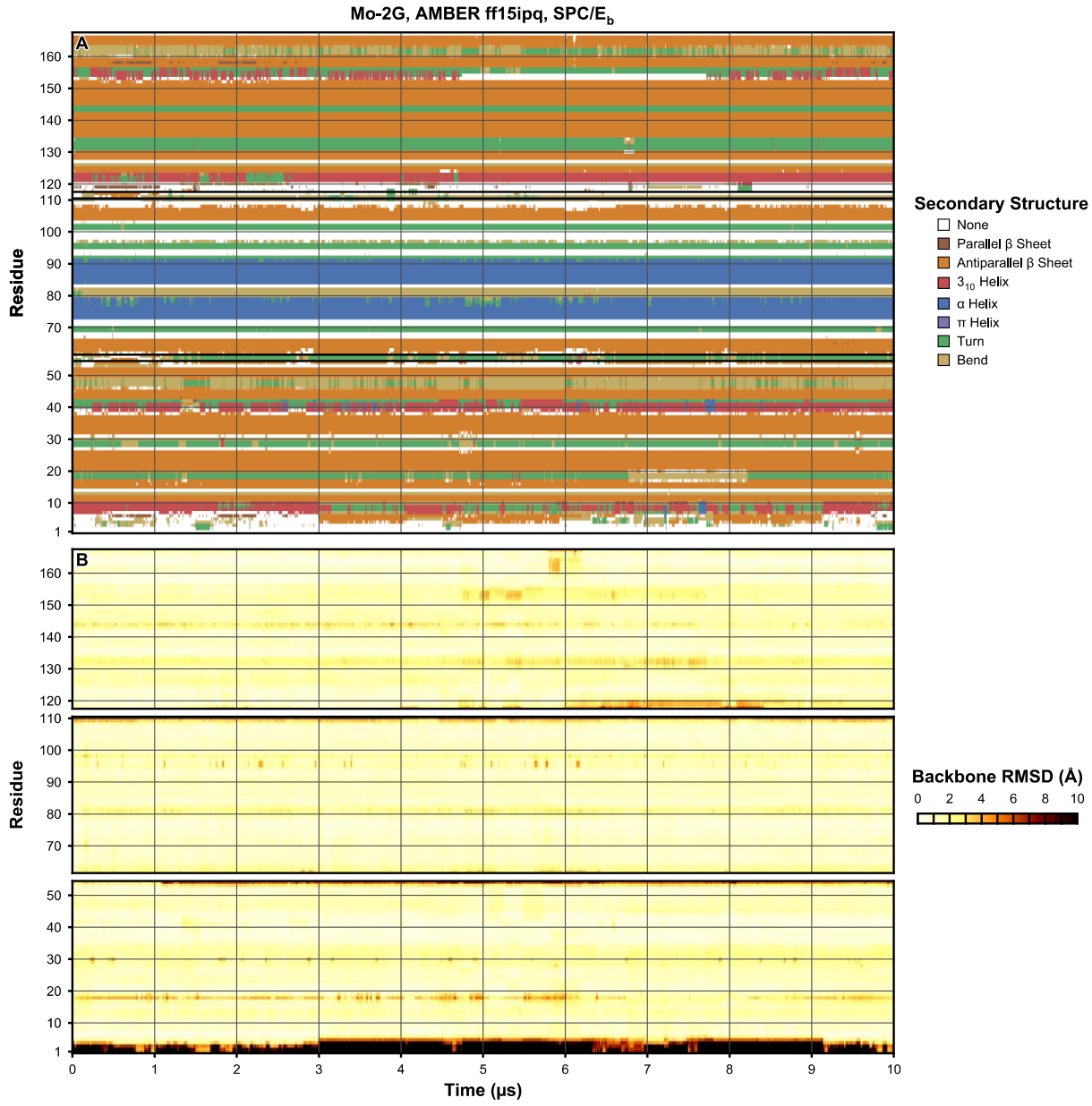

Figure 4.14. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the initial model (B) of Mo-2G over the course of a 10-μs simulation with f15ipq/SPC/E_b.

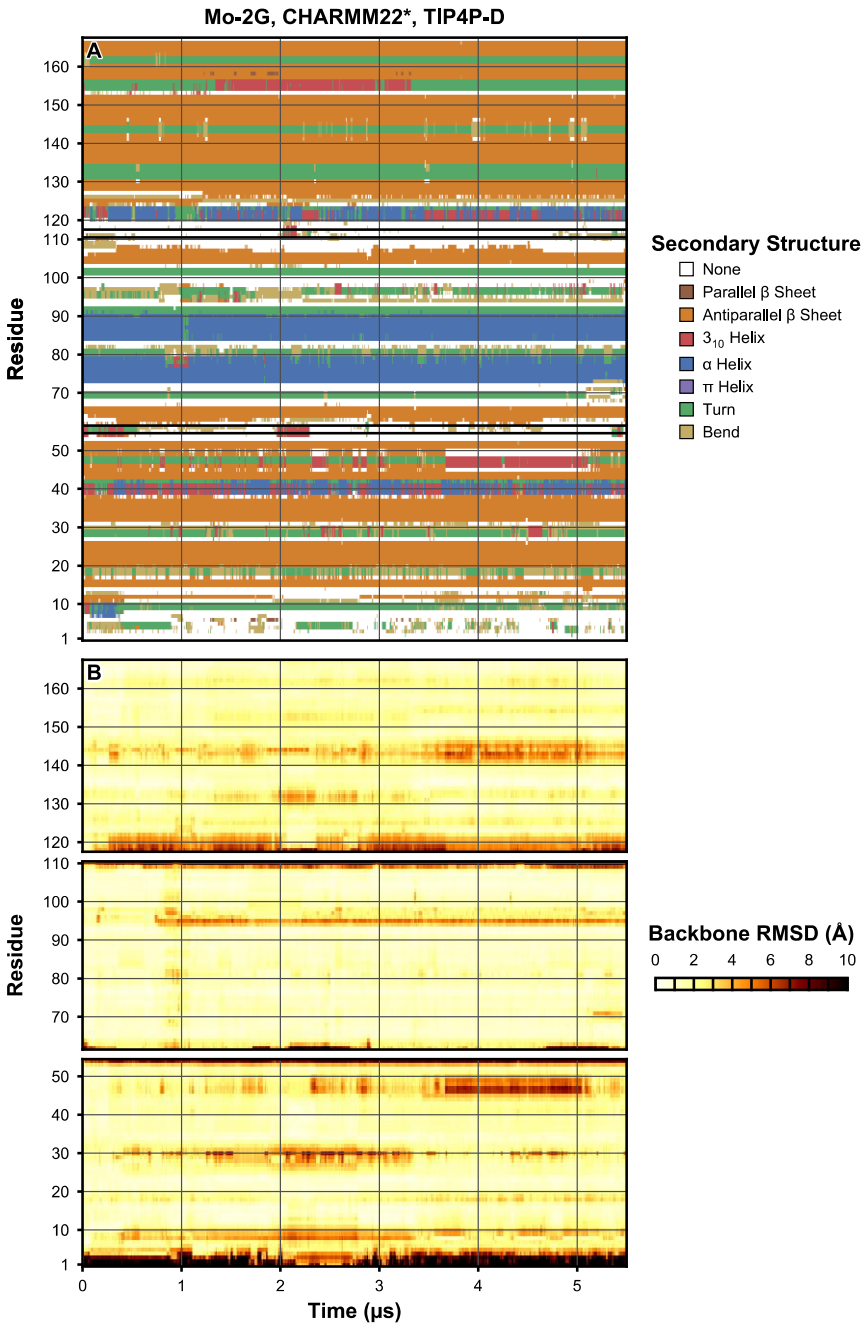

Figure 4.15. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the initial model (B) of Mo-2G over the course of a 5.5-μs simulation with CHARMM22*/TIP4P-D.

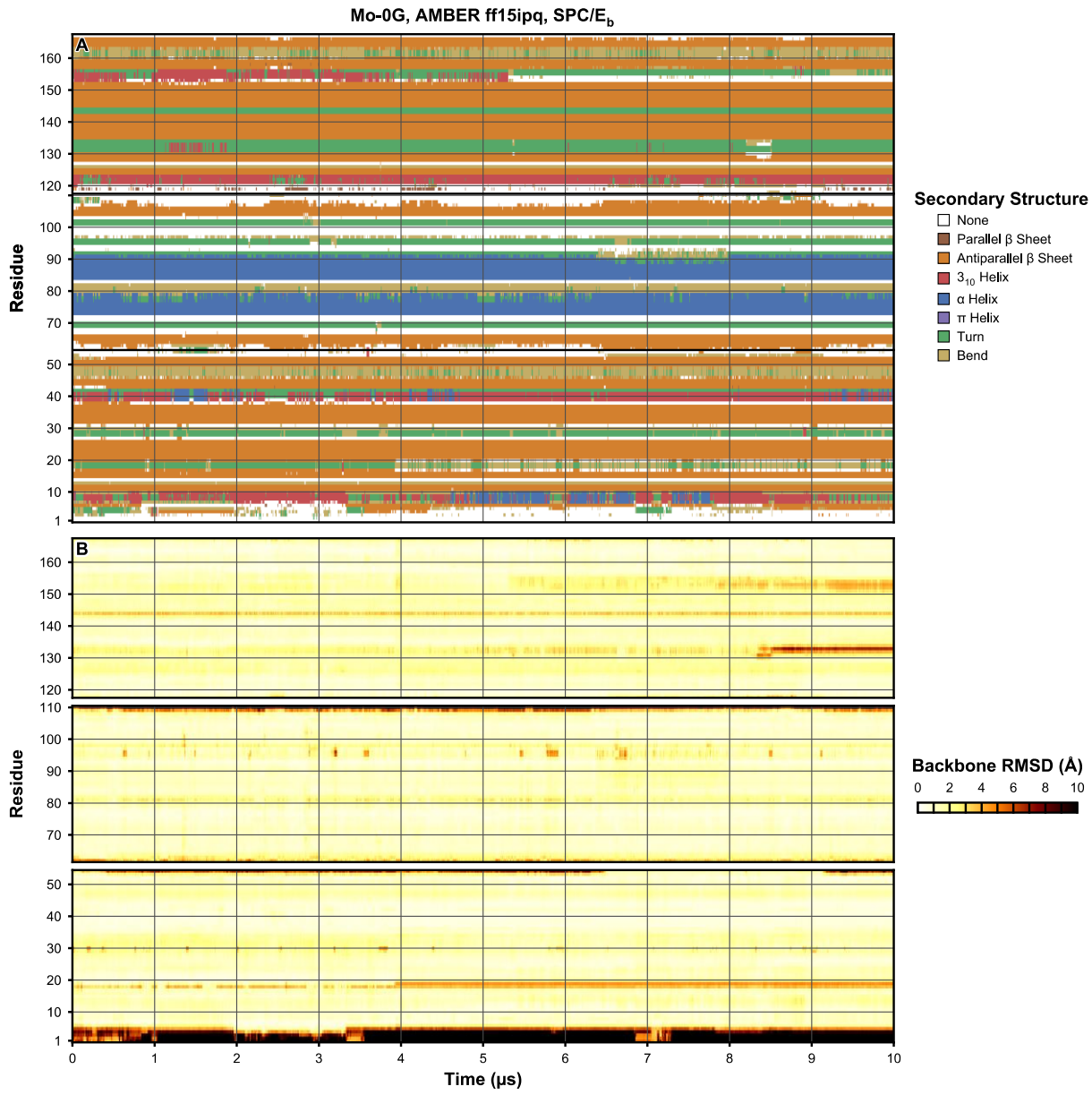

Figure 4.16. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the initial model (B) of Mo-0G over the course of a 10-μs simulation with f15ipq/SPC/E_b.

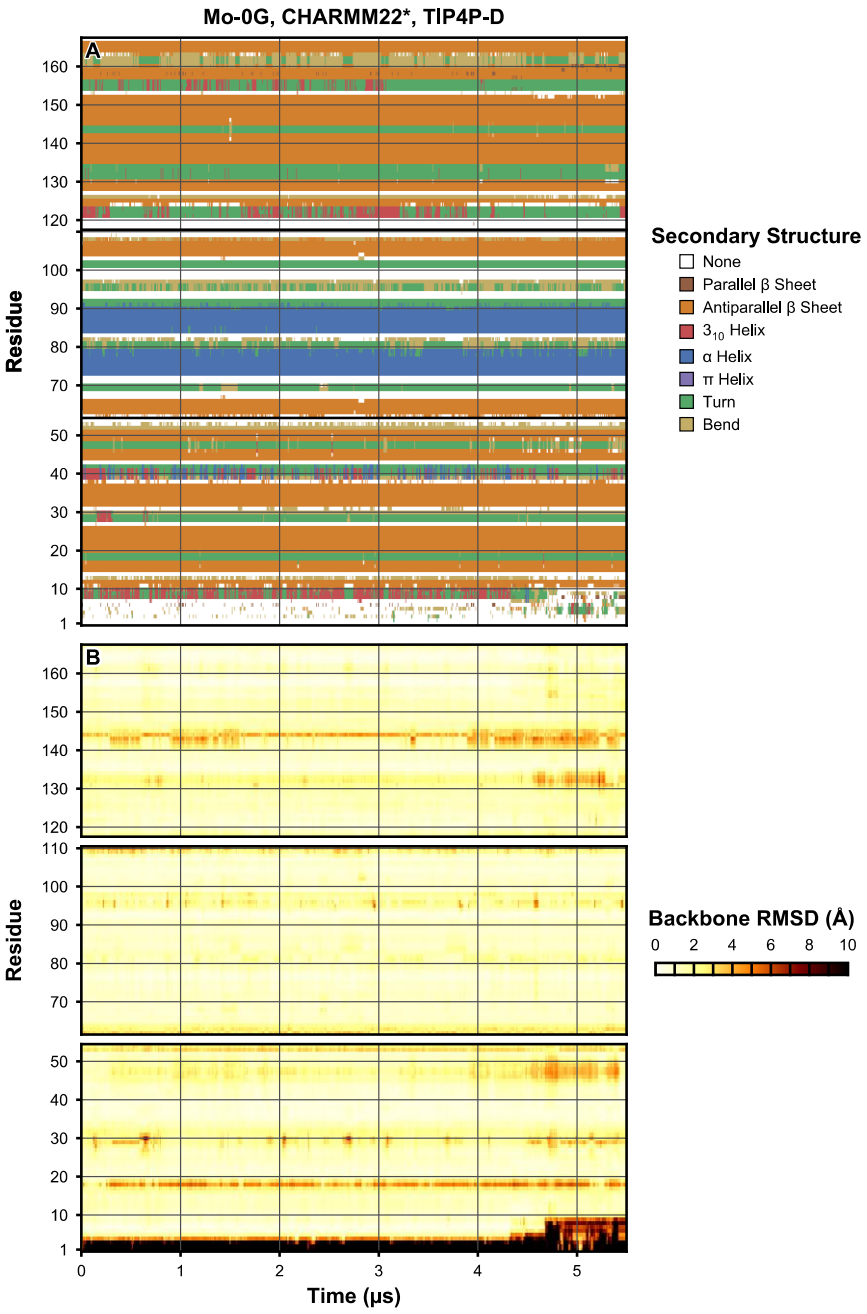

Figure 4.17. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the initial model (B) of Mo-0G over the course of a 5.5-μs simulation with CHARMM22*/TIP4P-D.

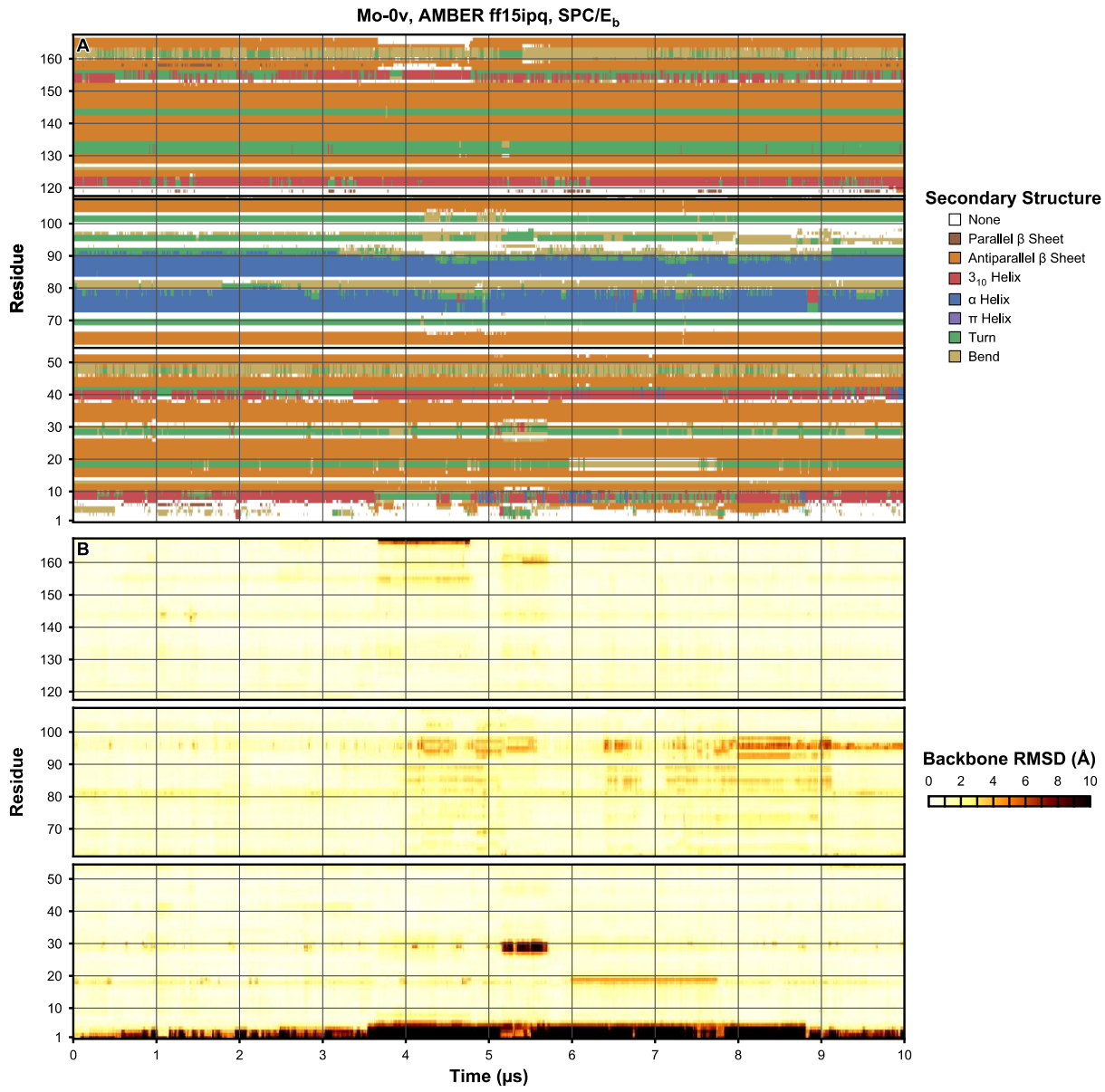

Figure 4.18. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the crystal structure (PDB code 5C8O) (B) of Mo-0v over the course of a 10-μs simulation with f15ipq/SPC/E_b.

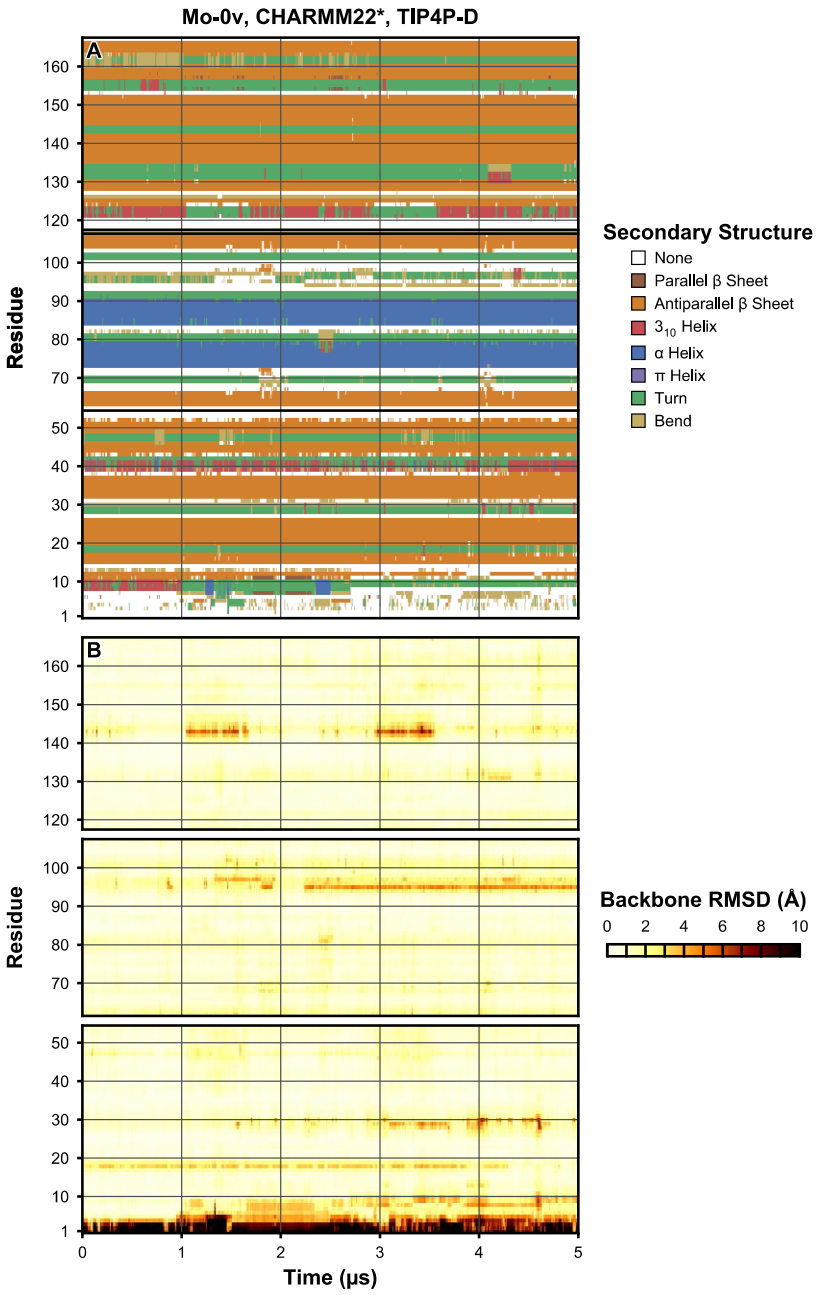

Figure 4.19. Secondary structure (A) and per-residue backbone RMSD for CVNH and LysM domain residues relative to the crystal structure (PDB code 5C8O) (B) of Mo-0v over the course of a 5-μs simulation with CHARMM22*/TIP4P-D.

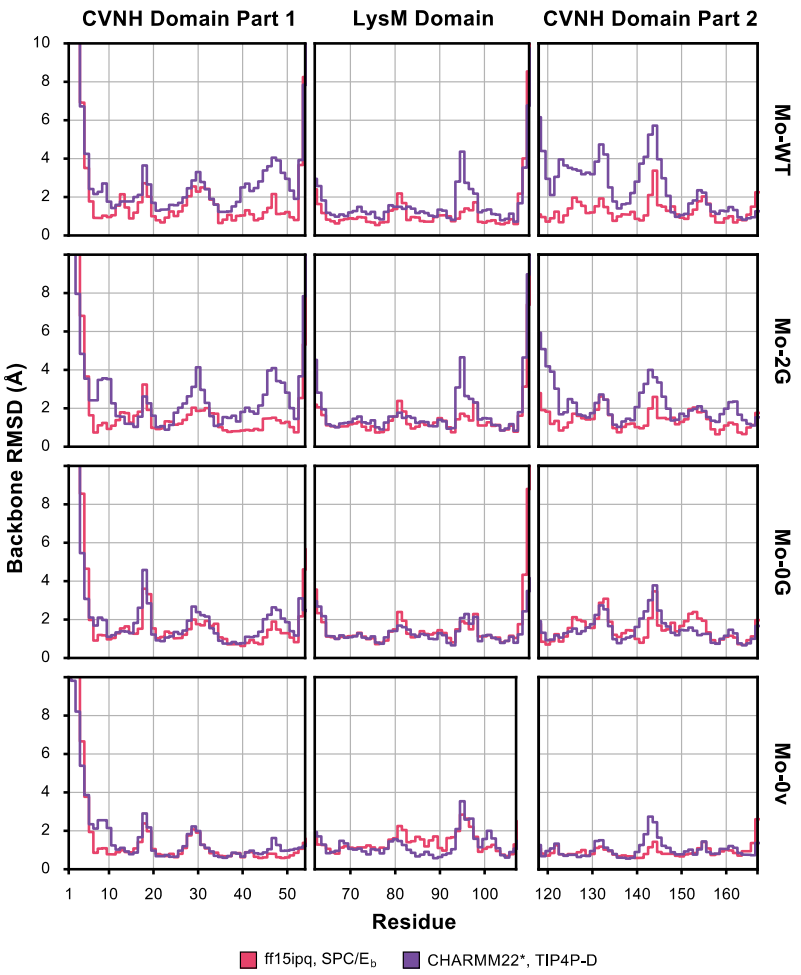

Figure 4.20. Distribution of per-residue backbone RMSD values of Mo-WT, Mo-2G, Mo-0G, and Mo0v relative to the initial structures over the course of simulations with ff15ipq/SPC/E_b (magenta) and CHARMM22*/TIP4P-D (purple). Lines represent the average RMSD values over the simulations, and shaded regions comprise the range between the 5th and 95th percentiles of sampled values.

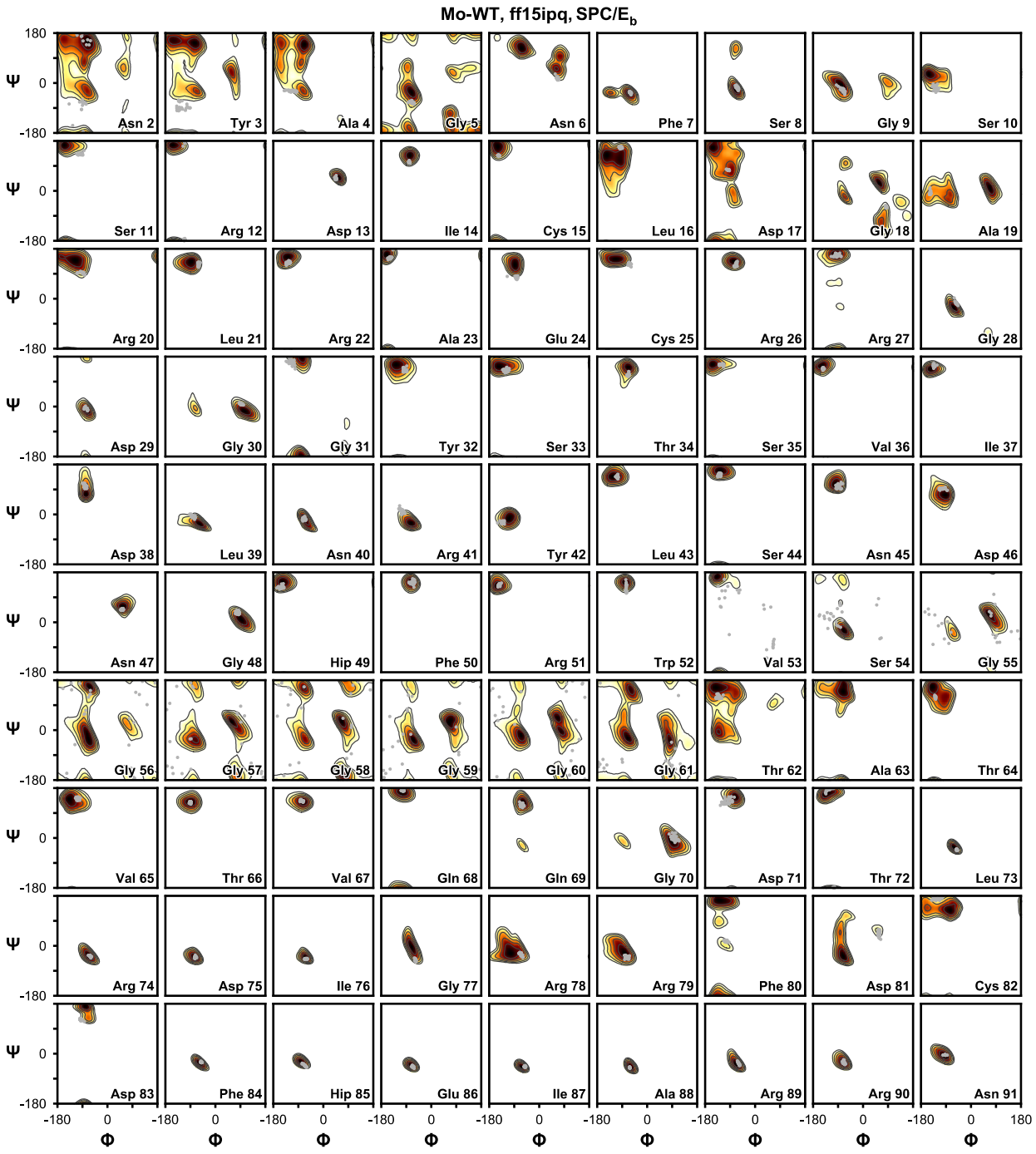

Figure 4.21. Backbone Φ/Ψ sampling for individual residues of Mo-WT over the course of a 10-μs simulation with ff15ipq/SPC/E_b. The corresponding Φ/Ψ angles in the NMR ensemble (PDB code 2L9Y) are shown in gray.

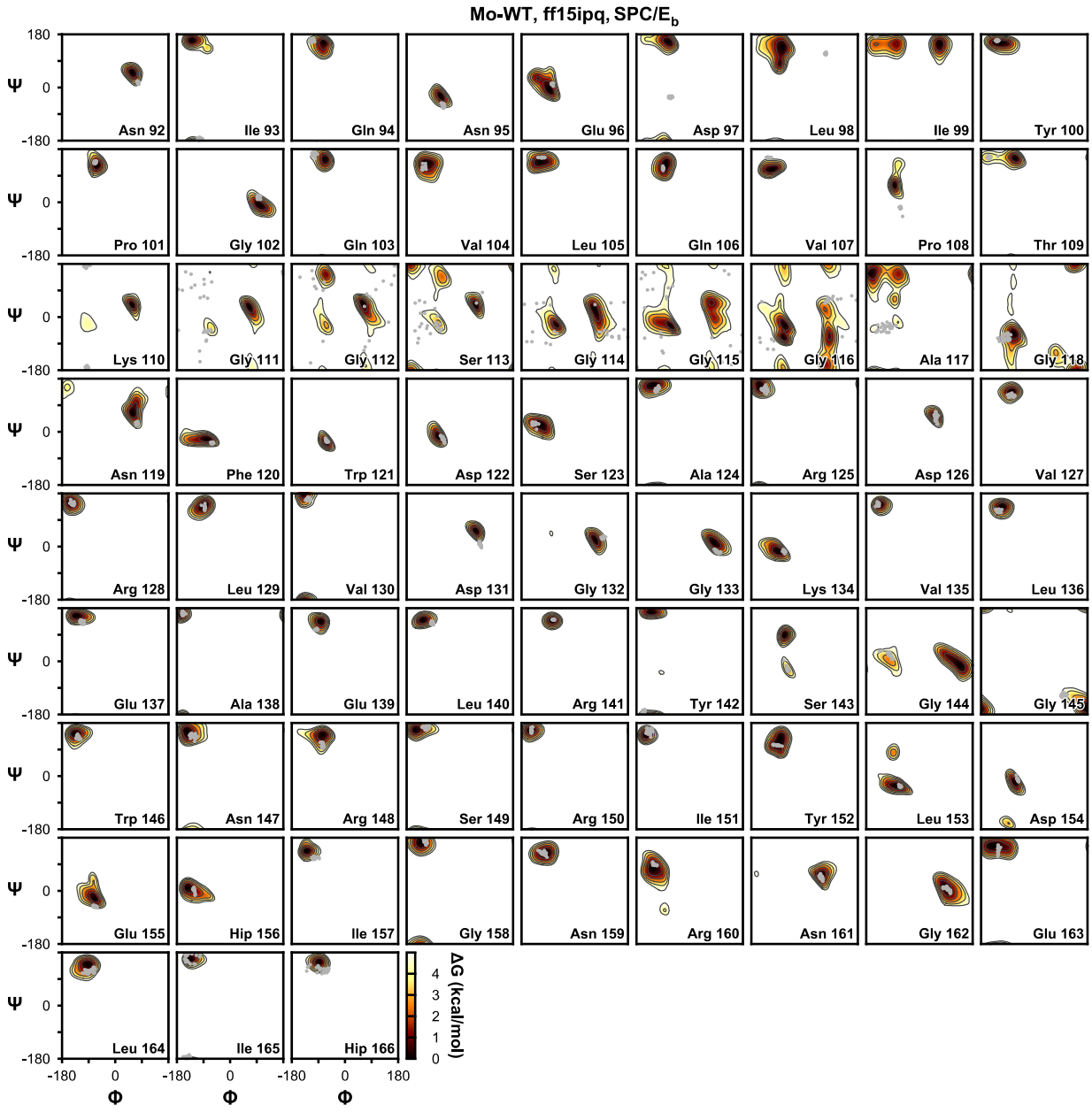

Figure 4.21 (Continued).

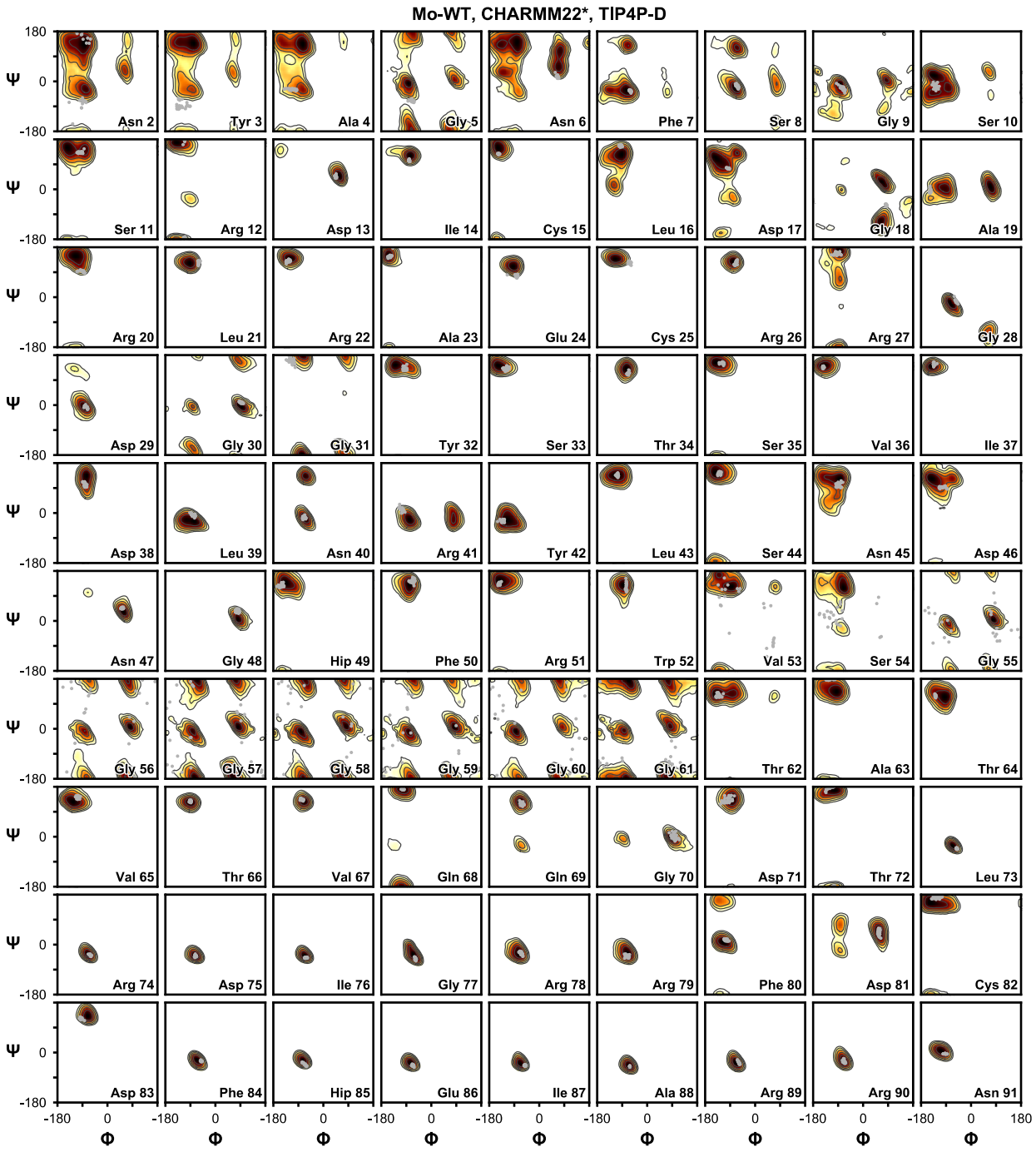

Figure 4.22. Backbone Φ/Ψ sampling for individual residues of Mo-WT over the course of a 7.3-μs simulation with CHARMM22*/TIP4P-D. The corresponding Φ/Ψ angles in the NMR ensemble (PDB code 2L9Y) are shown in gray.

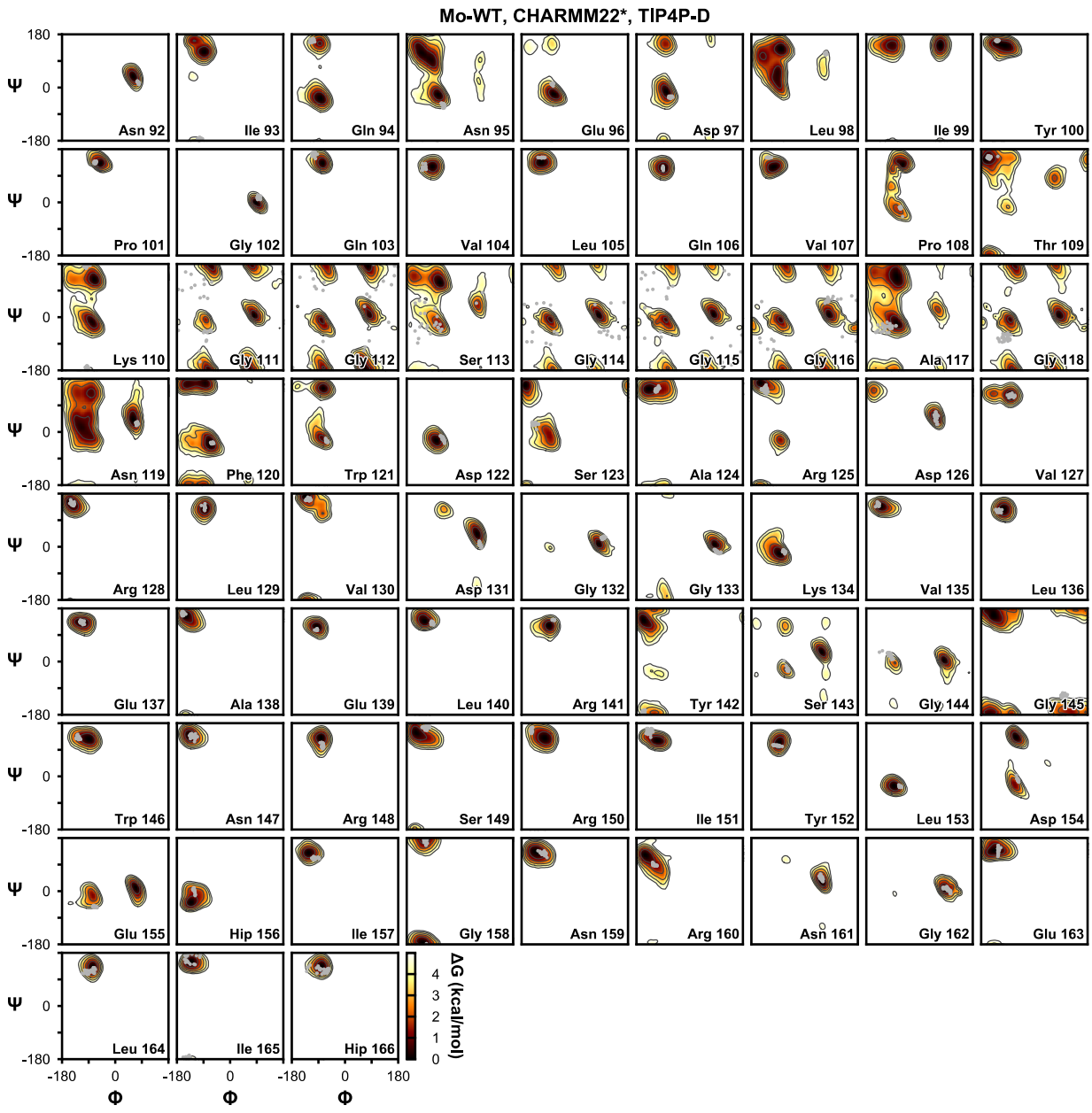

Figure 4.22 (Continued).

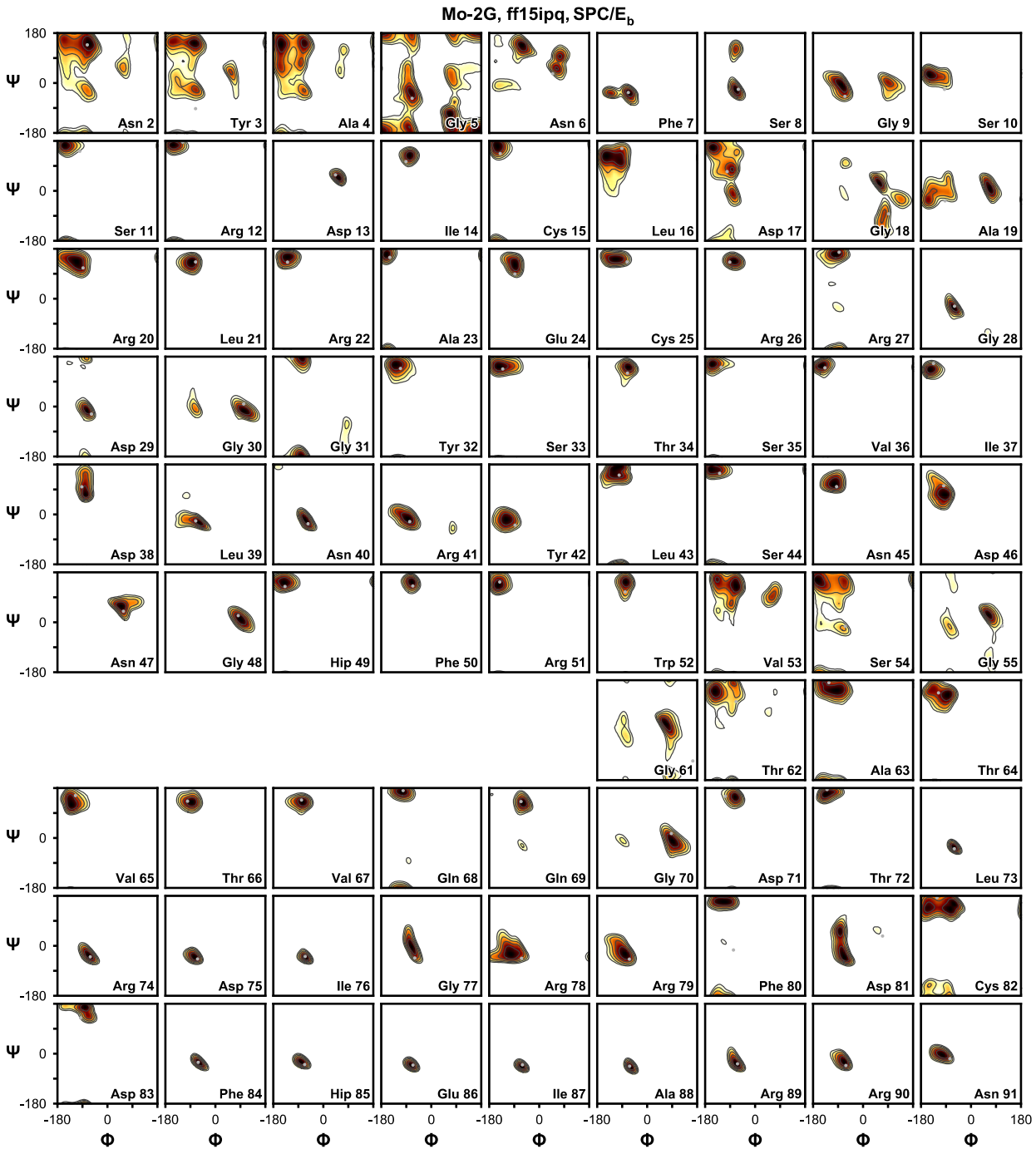

Figure 4.23. Backbone Φ/Ψ sampling for individual residues of Mo-2G over the course of a 10-μs simulation with ff15ipq/SPC/E_b. The corresponding Φ/Ψ angles in the initial model are shown in gray.

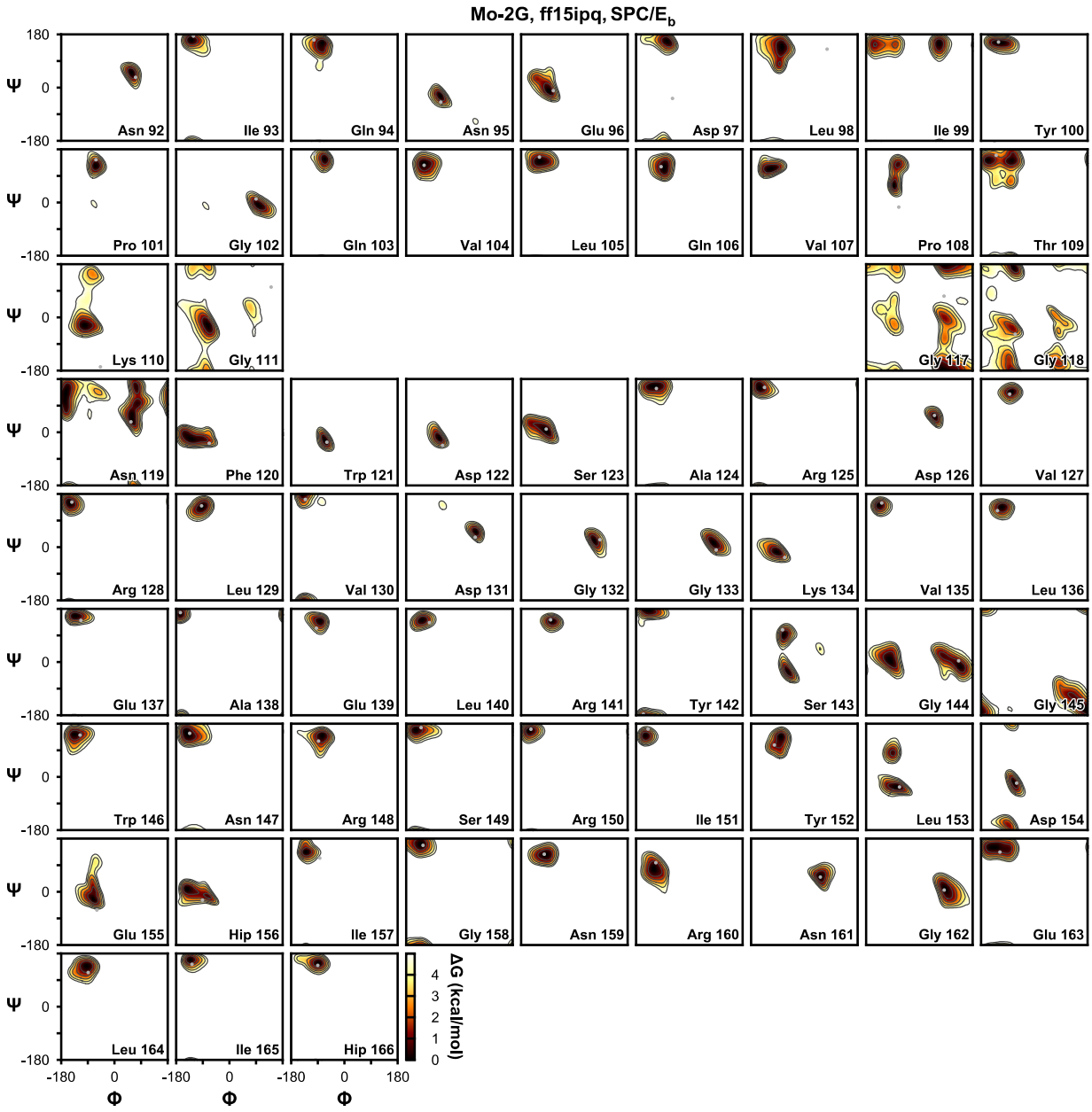

Figure 4.23 (Continued).

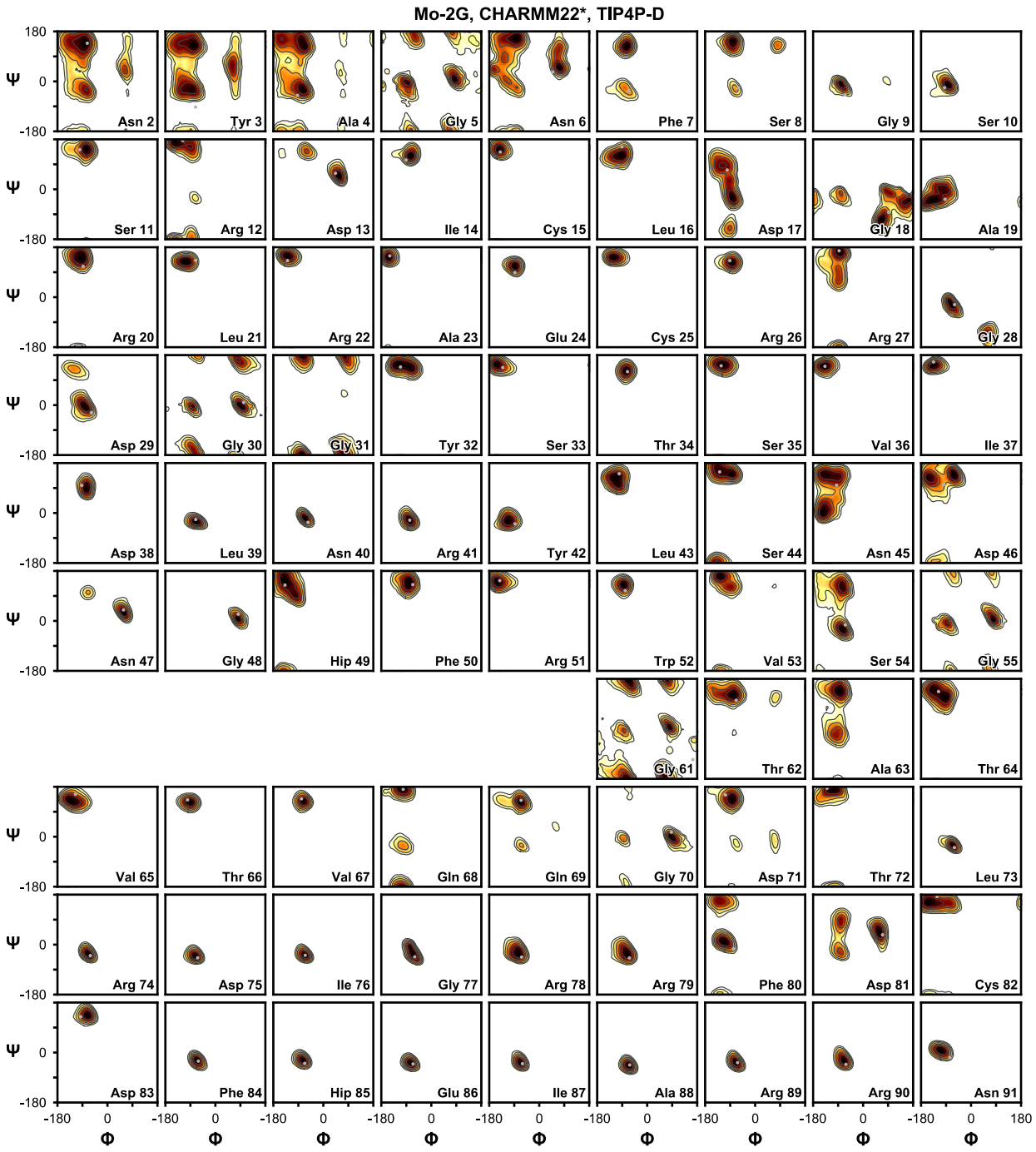

Figure 4.24. Backbone Φ/Ψ sampling for individual residues of Mo-2G over the course of a 5.5-μs simulation with CHARMM22*/TIP4P-D. The corresponding Φ/Ψ angles in the initial model are shown in gray.

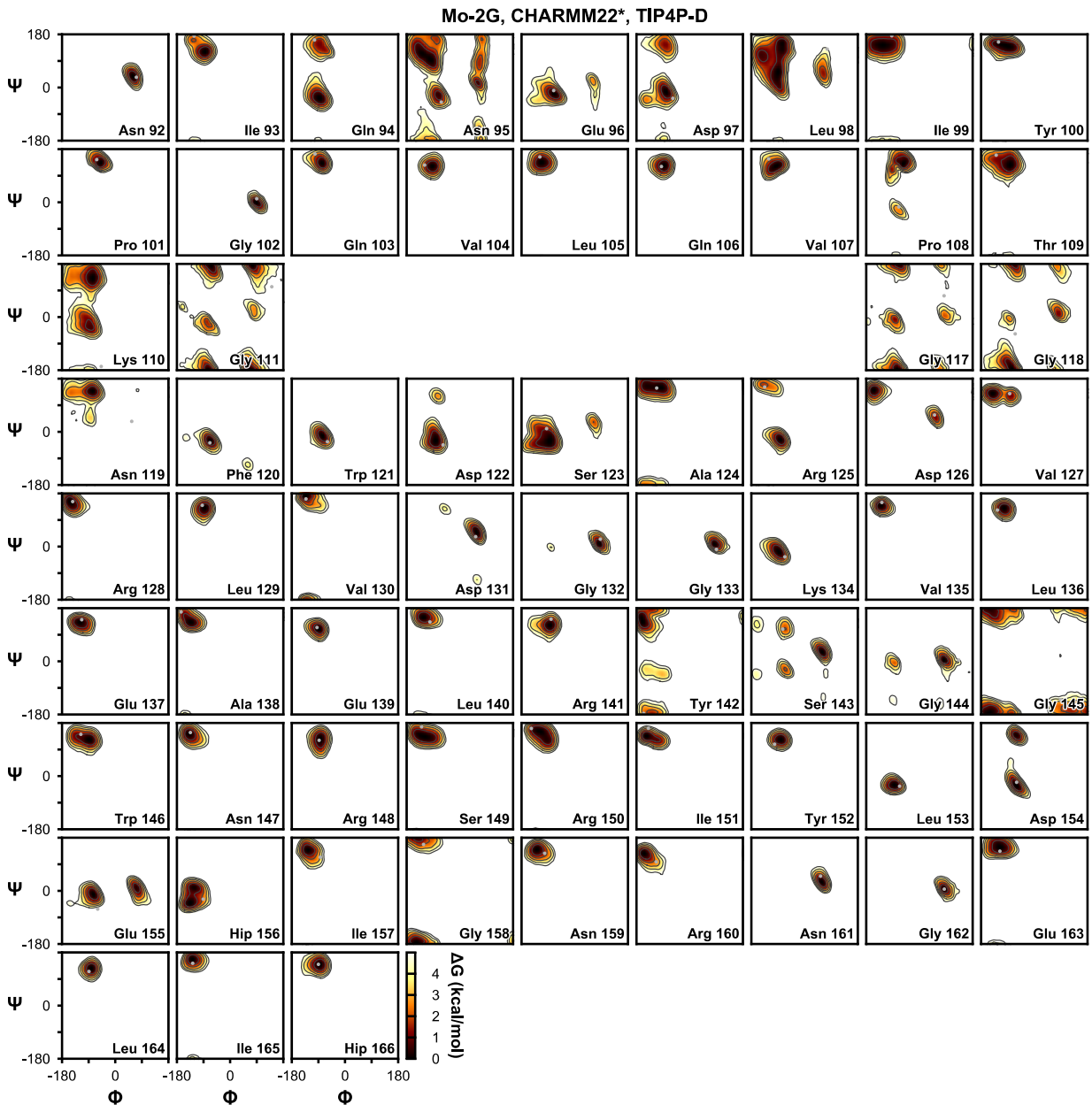

Figure 4.24 (Continued).

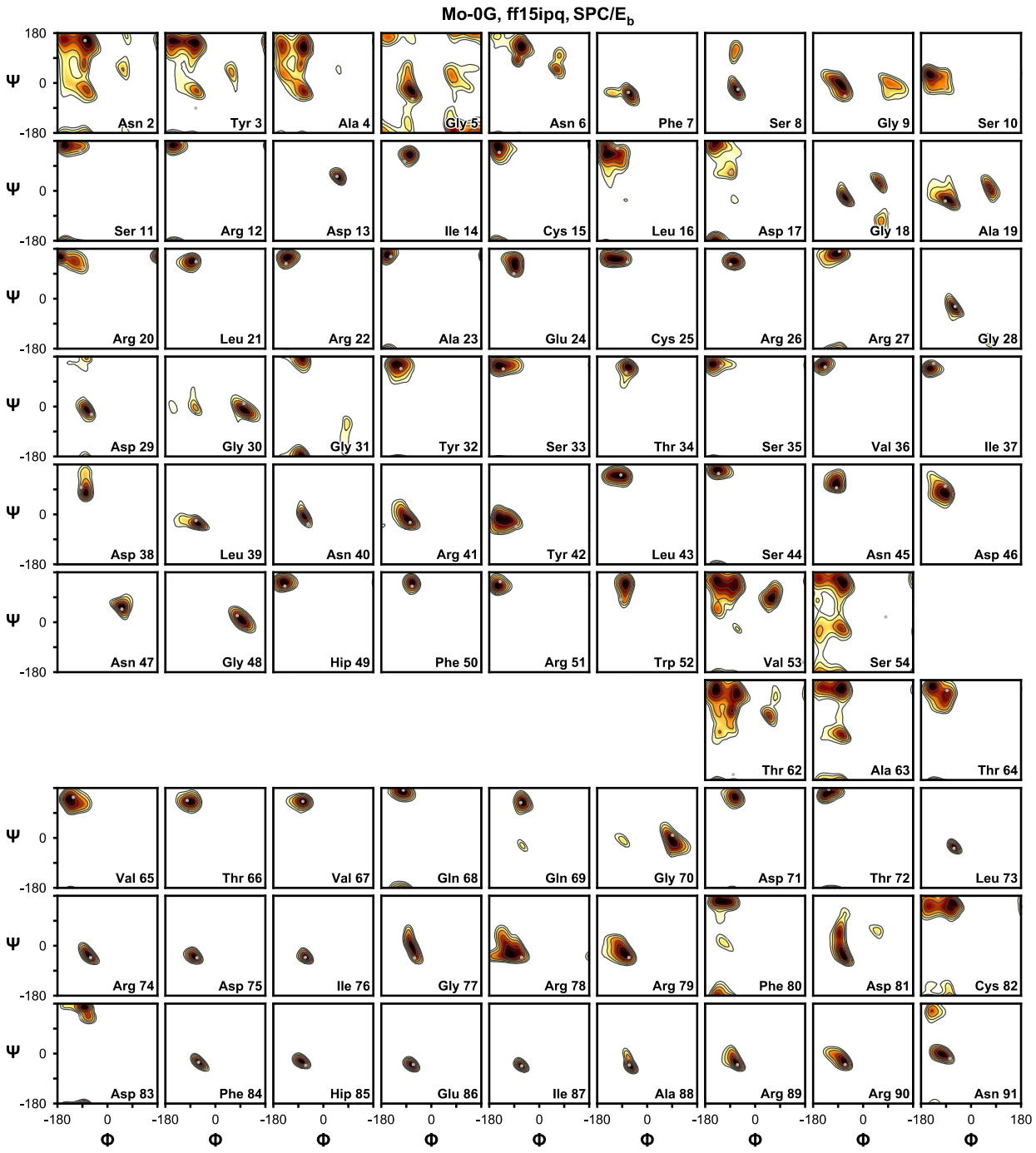

Figure 4.25. Backbone Φ/Ψ sampling for individual residues of Mo-0G over the course of a 10-μs simulation with ff15ipq/SPC/E_b. The corresponding Φ/Ψ angles in the initial model are shown in gray.

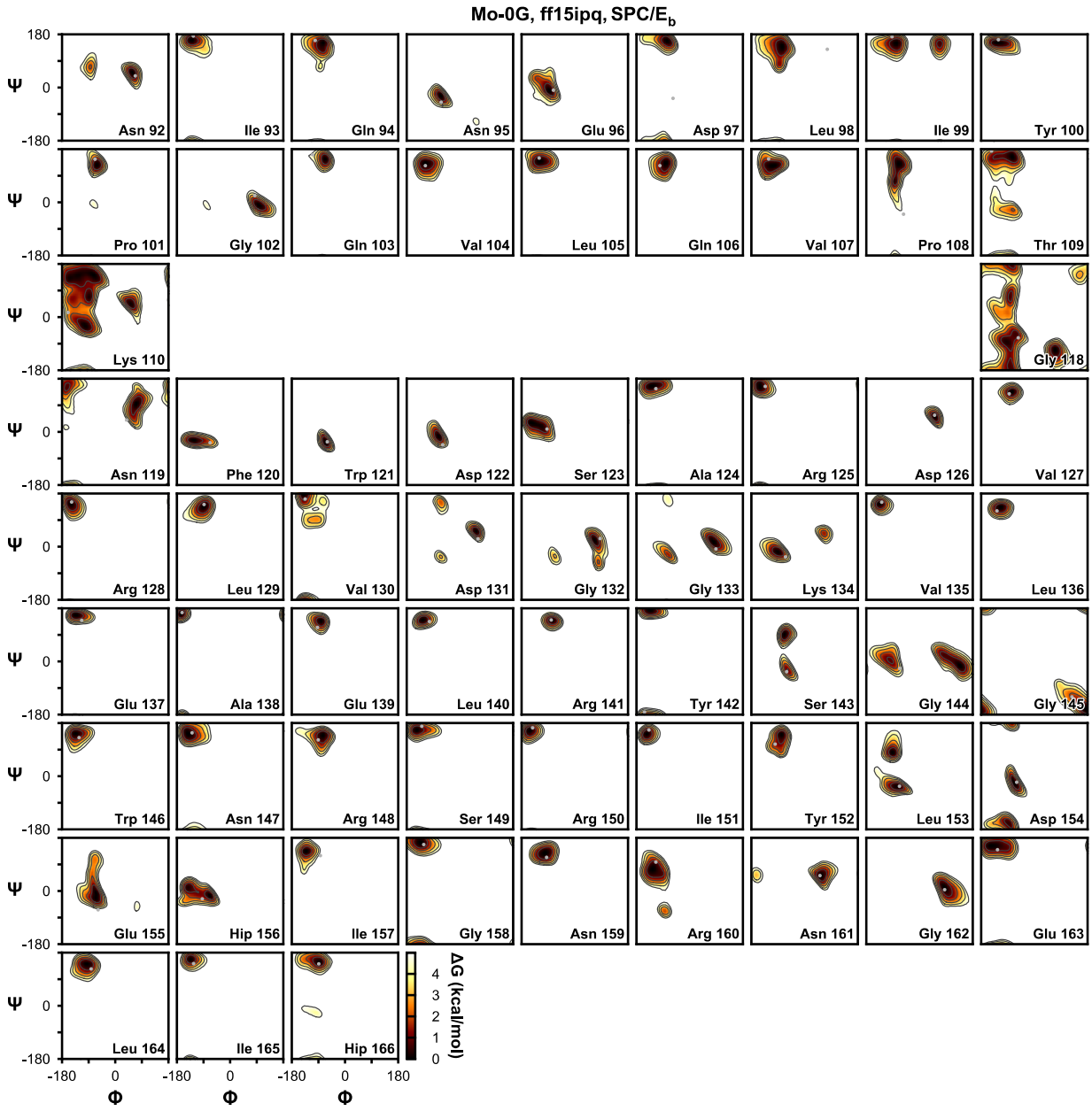

Figure 4.25 (Continued).

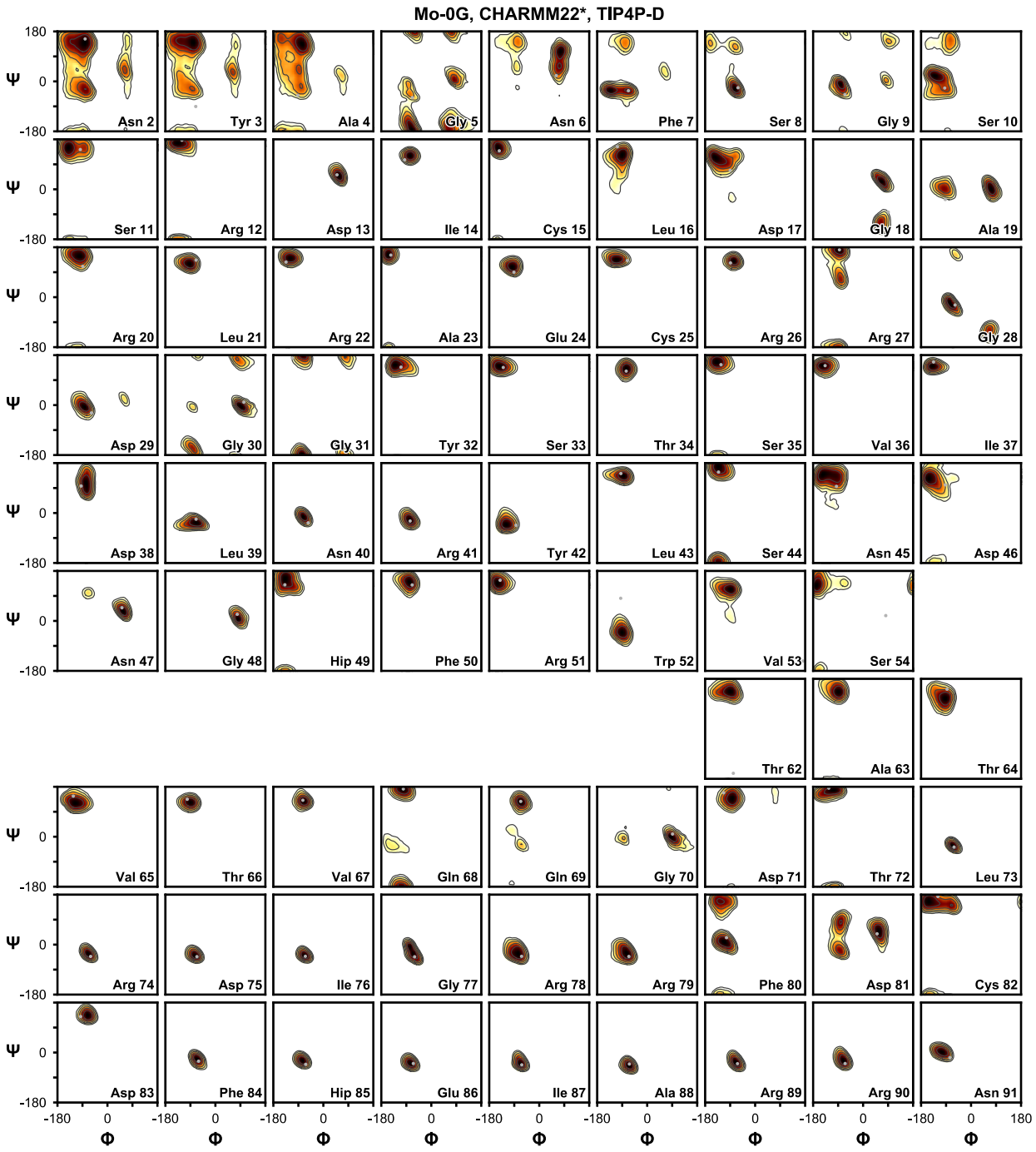

Figure 4.26. Backbone Φ/Ψ sampling for individual residues of Mo-0G over the course of a 5.5-μs simulation with CHARMM22*/TIP4P-D. The corresponding Φ/Ψ angles in the initial model are shown in gray.

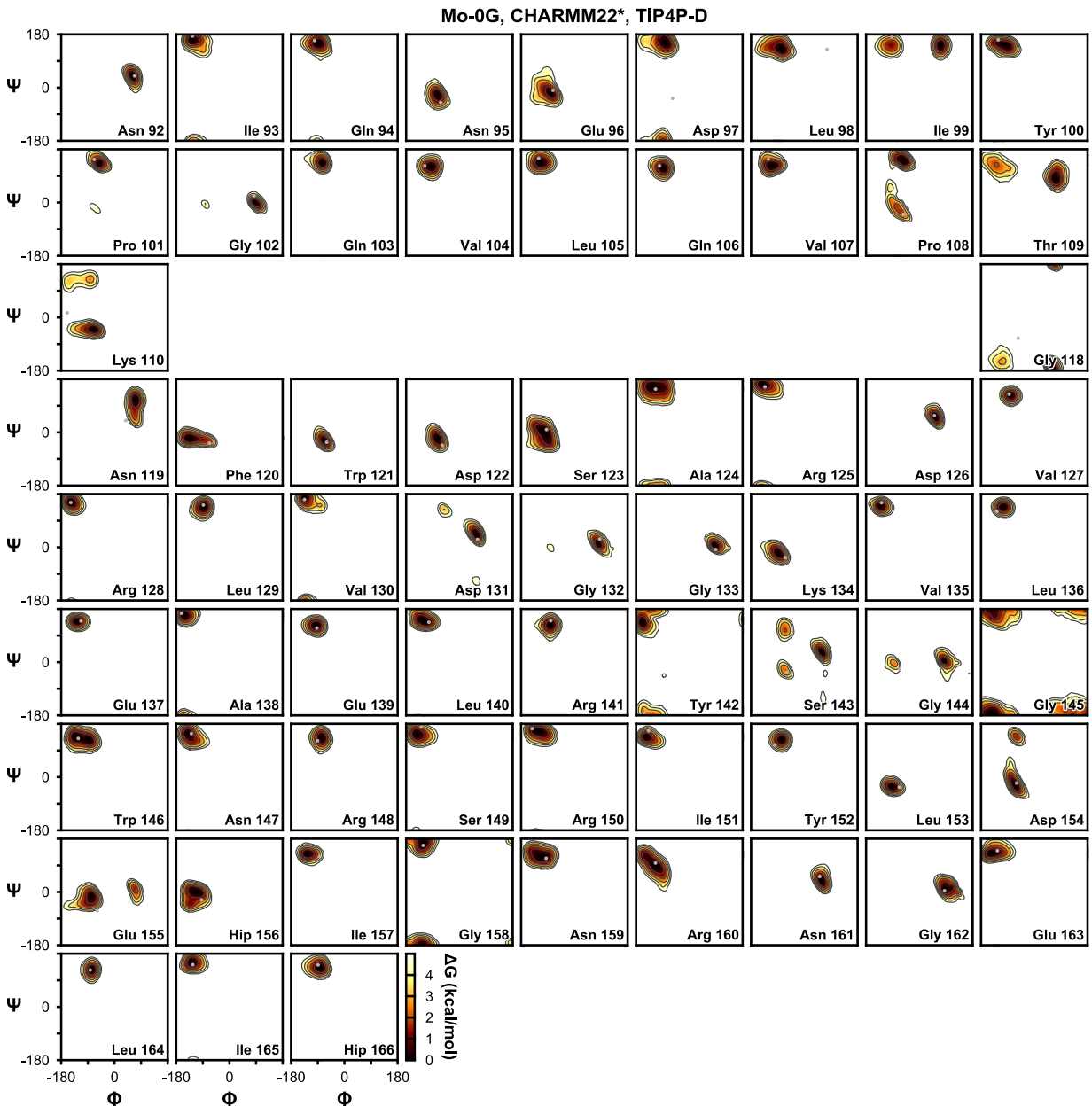

Figure 4.26 (Continued).

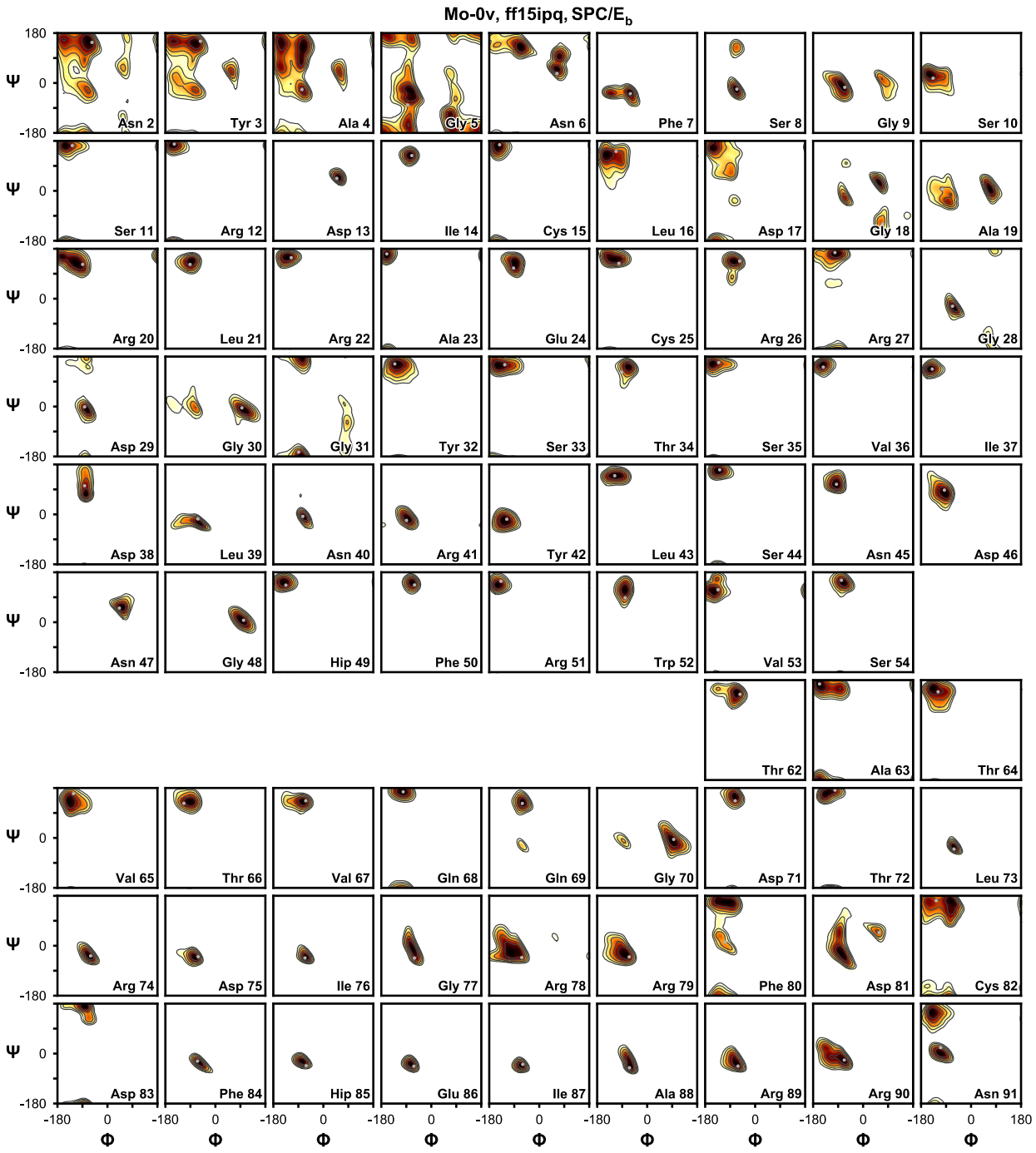

Figure 4.27. Backbone Φ/Ψ sampling for individual residues of Mo-0v over the course of a 10-μs simulation with ff15ipq/SPC/E_b. The corresponding Φ/Ψ angles in the crystal structure (PDB code 5C8O) are shown in gray.

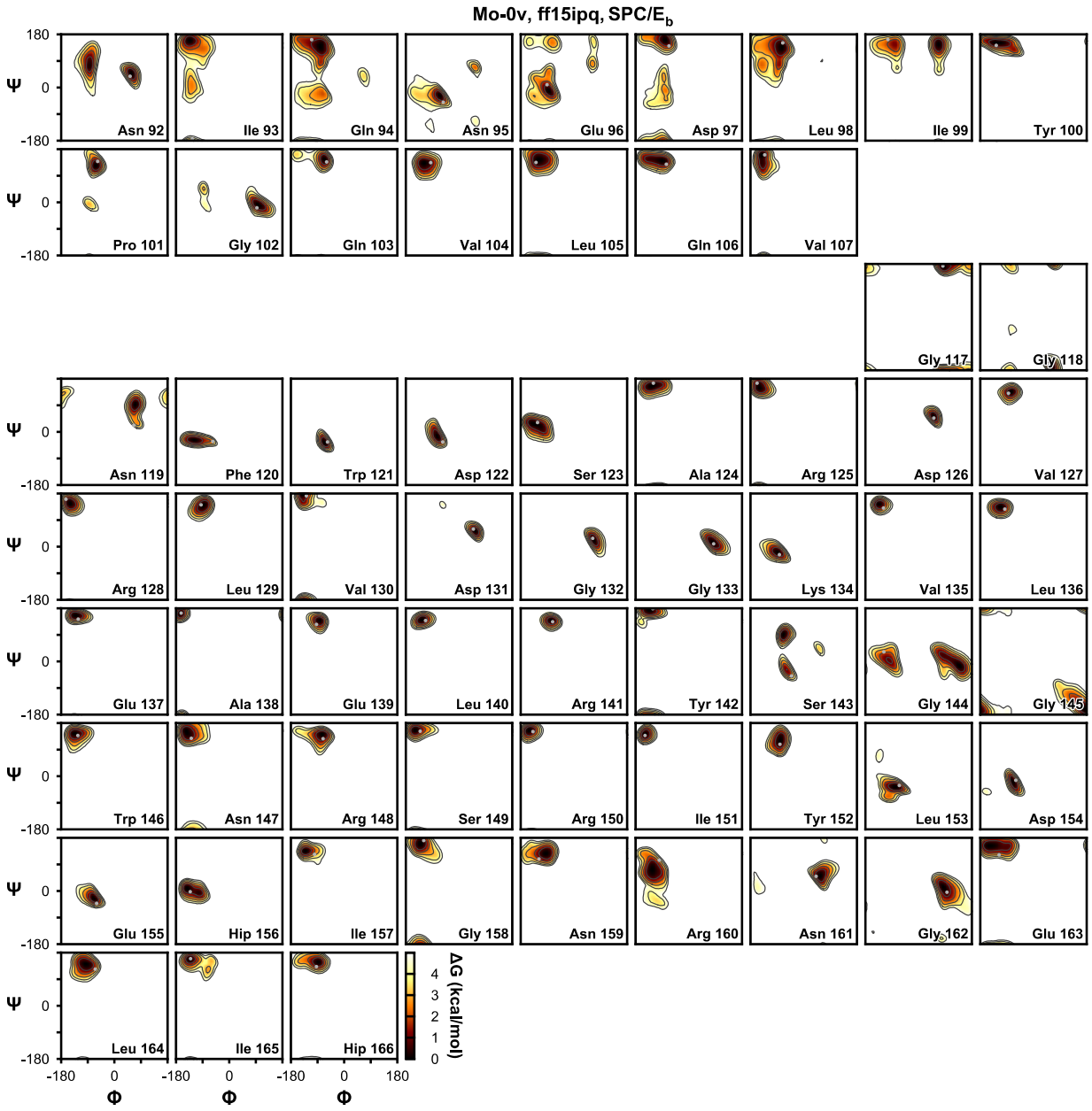

Figure 4.27 (Continued).

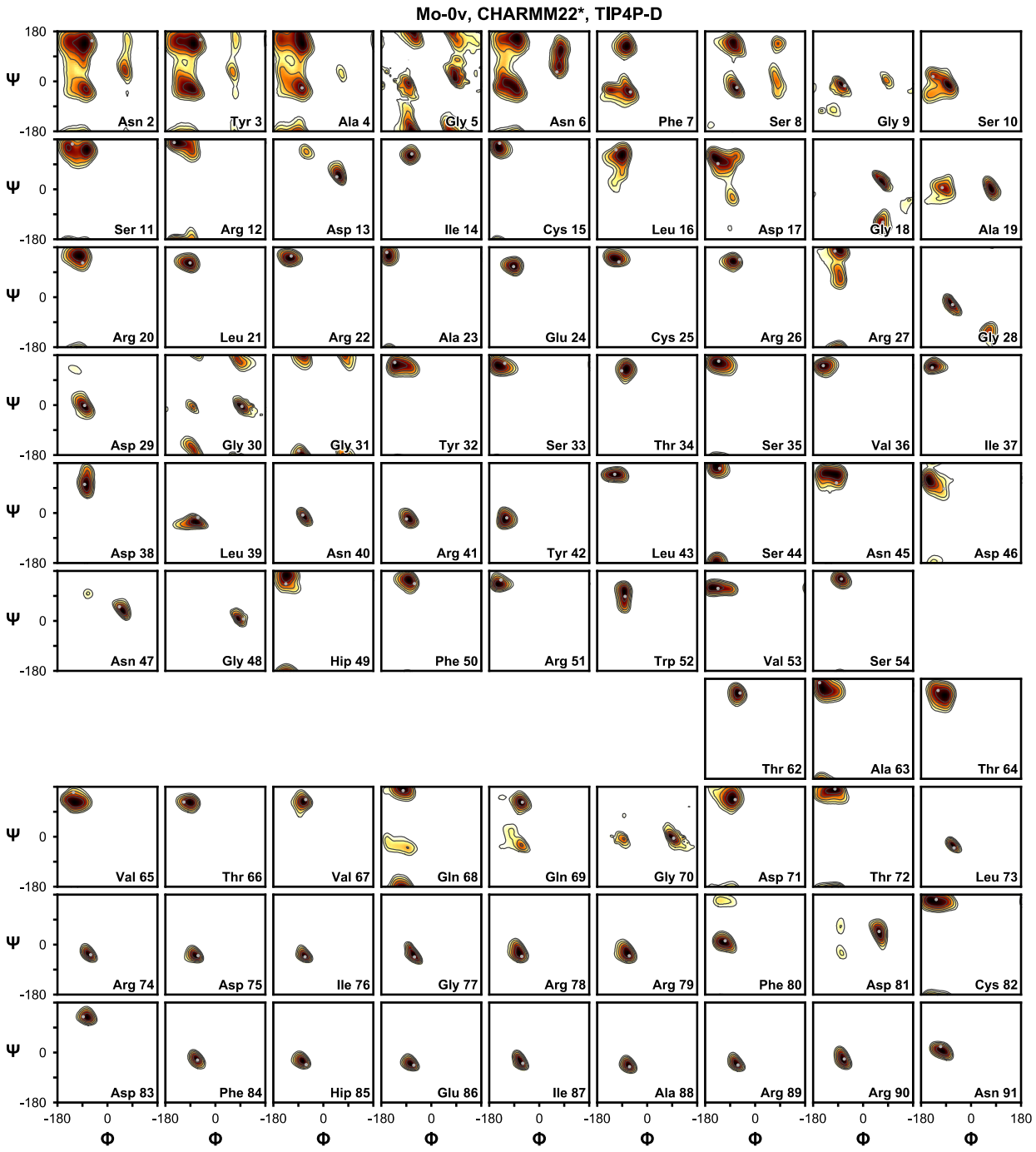

Figure 4.28. Backbone Φ/Ψ sampling for individual residues of Mo-0v over the course of a 5-μs simulation with CHARMM22*/TIP4P-D. The corresponding Φ/Ψ angles in the crystal structure (PDB code 5C8O) are shown in gray.

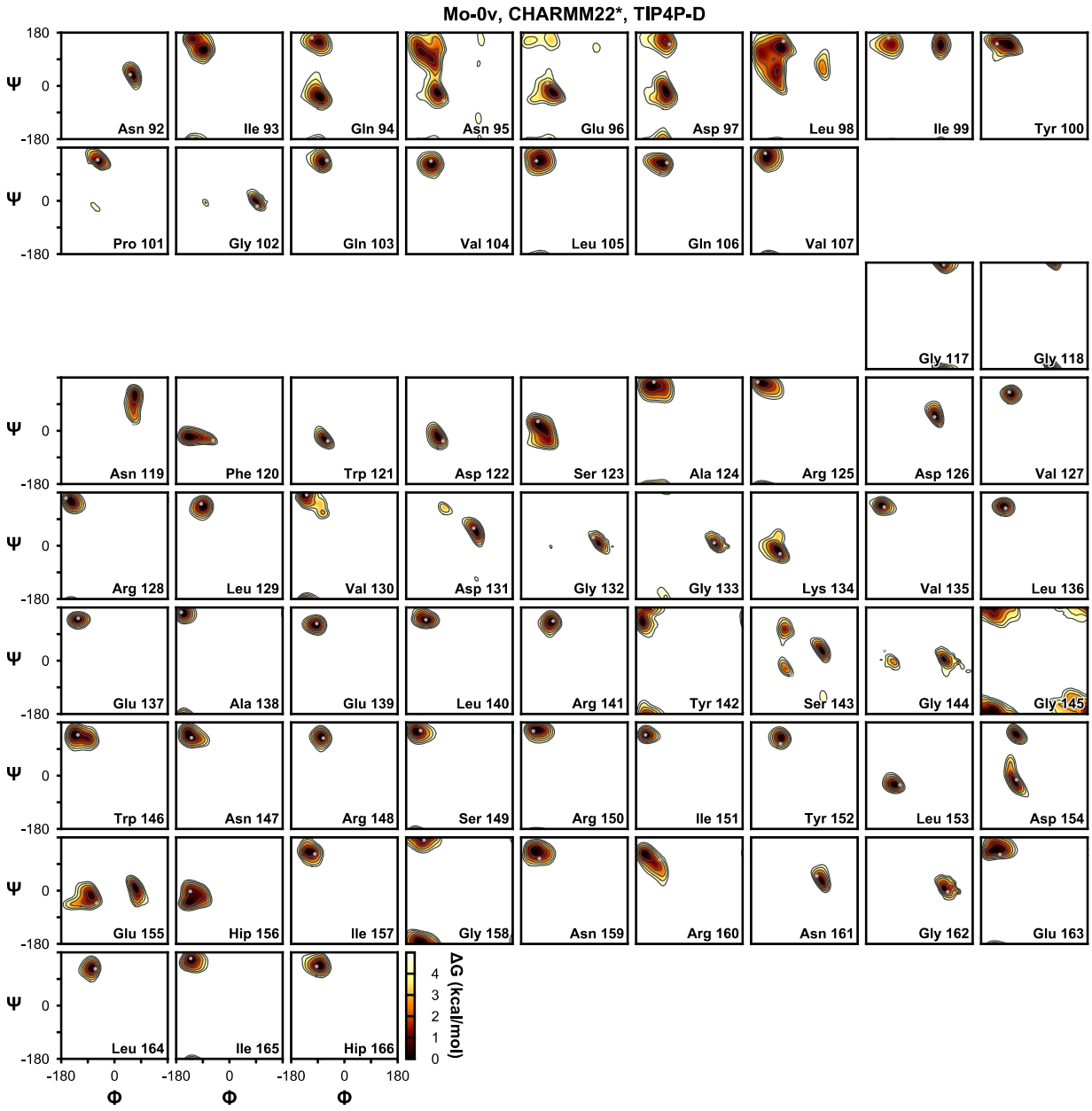

Figure 4.28 (Continued).

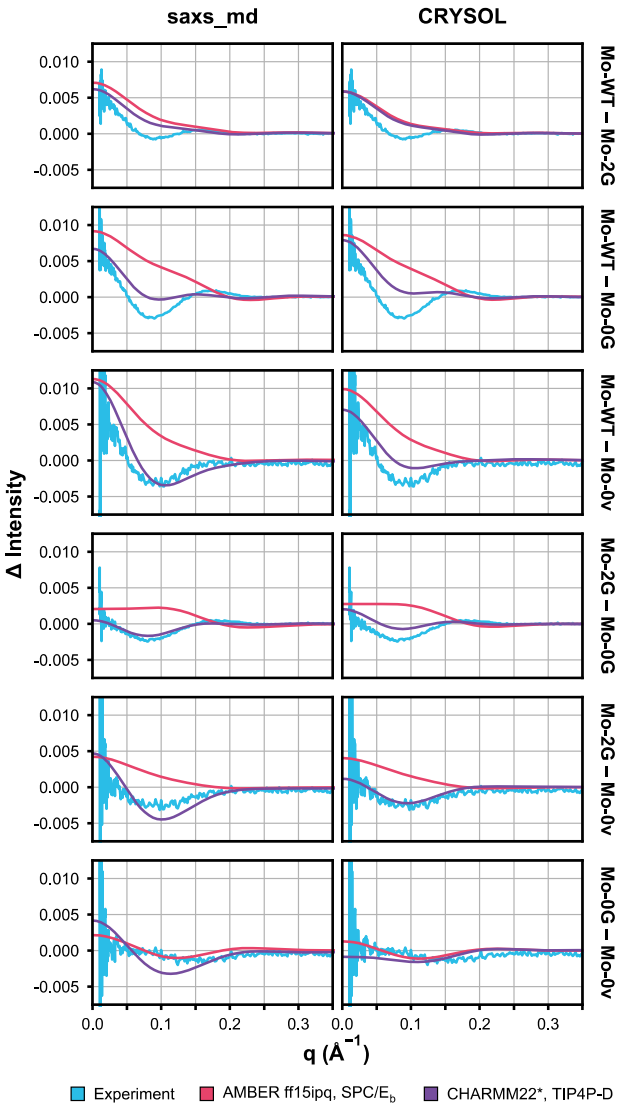

Figure 4.29. Difference in experimentally measured (cyan), and back-calculated (magenta, purple) small-angle X-ray scattering intensity between the different MoCVNH3 constructs. Y-axis units are arbitrary; shaded regions represent 95% confidence intervals of the experimental and simulated values.

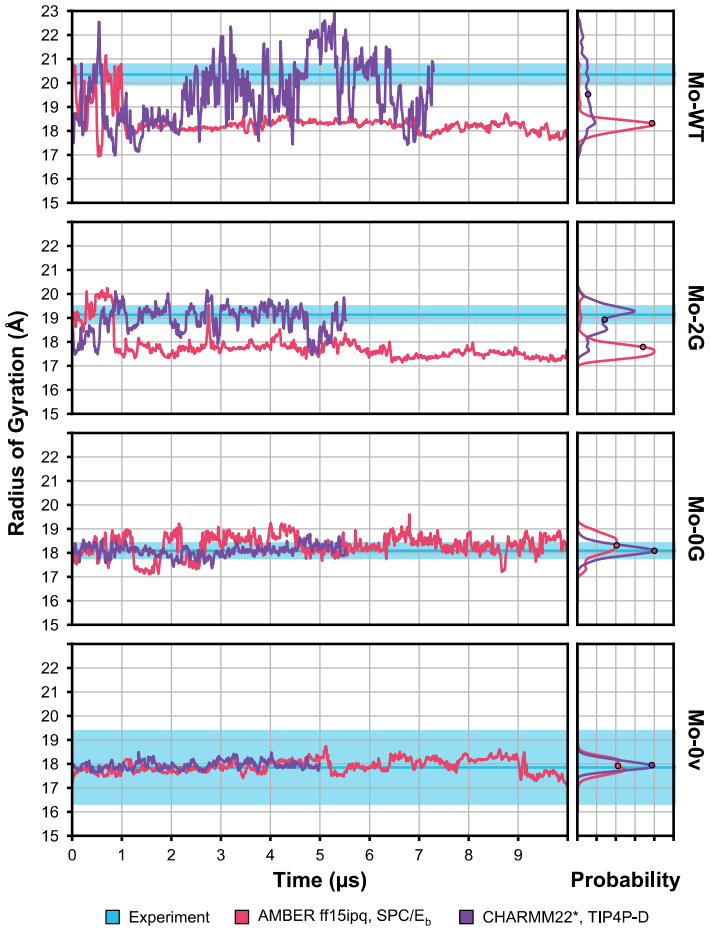

Figure 4.30. Radius of gyration (R_g) of MoCVNH3 constructs over the course of simulations with ff15ipq/SPC/E_b (magenta) and CHARMM22*/TIP4P-D (purple). The probability distributions of sampled values are shown in the right panels, with the average values marked by circles. Experimental radii of gyration calculated from the SAXS data are shown in cyan, and the shaded regions represent 95% confidence intervals on the experimental values.

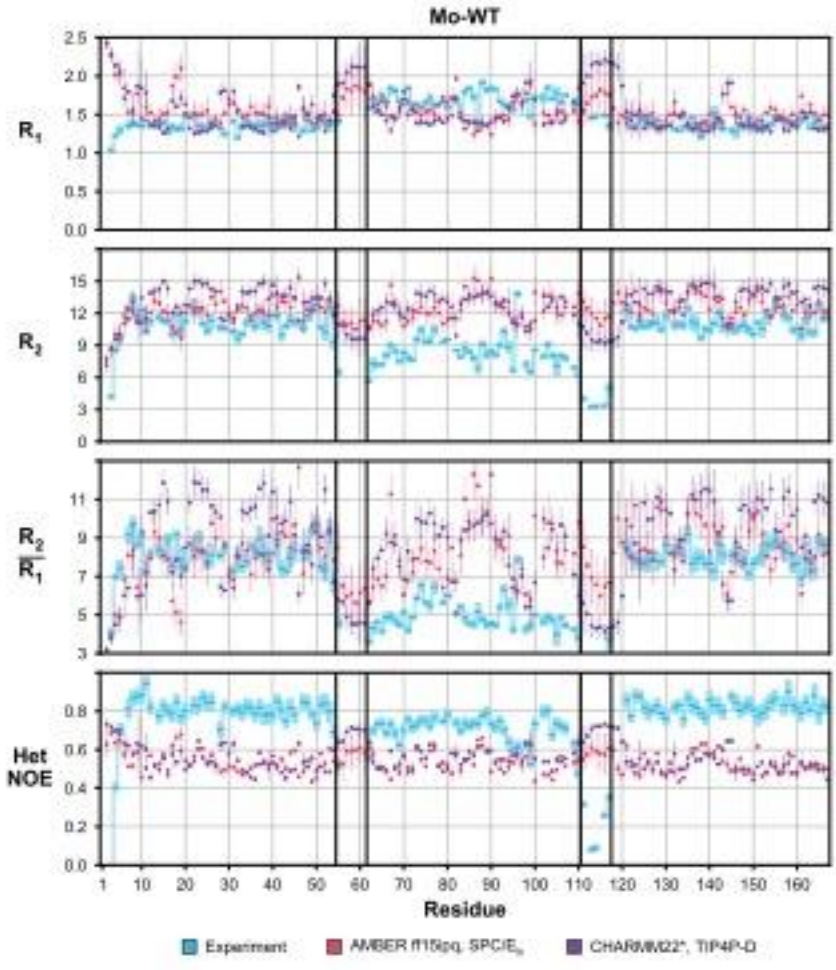

Figure 4.31. Experimental (cyan) and simulated ^{15}N R_1, R_2, and R_2/R_1 ratio and ^{1}H-^{15}N heteronuclear NMR relaxation data for Mo-WT from simulations with ff15ipq/SPC/E_b (magenta) and CHARMM22*/TIP4P-D (purple), calculated using a rolling 500-ns window. Shaded regions of the experimental values and the vertical bars of the simulated values represent 95% confidence intervals.

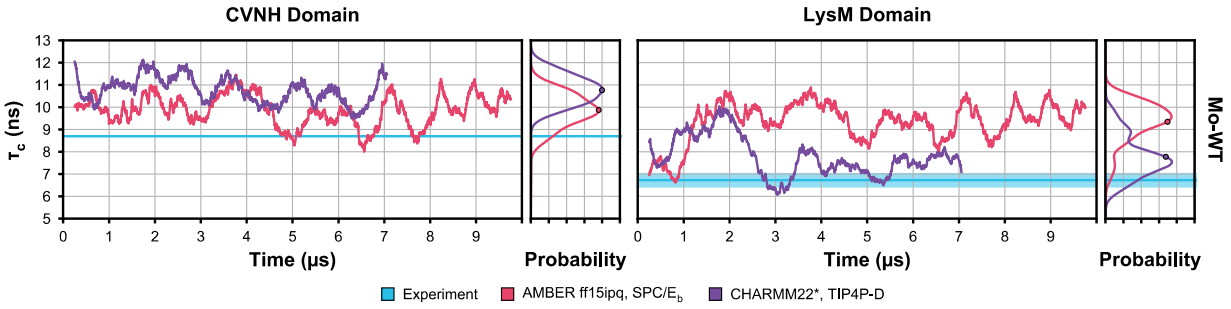

Figure 4.32. Rotational correlation time (τ_c) of the CVNH and LysM domains of Mo-WT over the course of simulations with ff15ipq/SPC/E_b (magenta) and CHARMM22*/TIP4P-D (purple), calculated using a rolling 500-ns window. The probability distribution of sampled values are shown in the right panels, with average values marked by circles. Experimental τ_c values calculated from the NMR relaxation data are shown in cyan, and the shaded regions represent 95% confidence intervals on the experimental values.

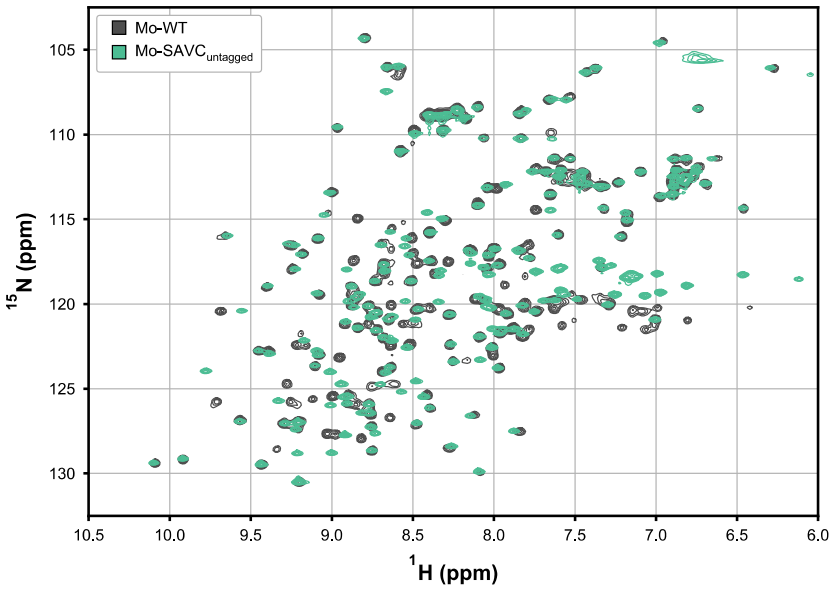

Figure 4.33. Superposition of ^{1}H-^{15}N HSQC spectra of Mo-WT (black) and the single-cysteine mutant Mo-SAVC (green).

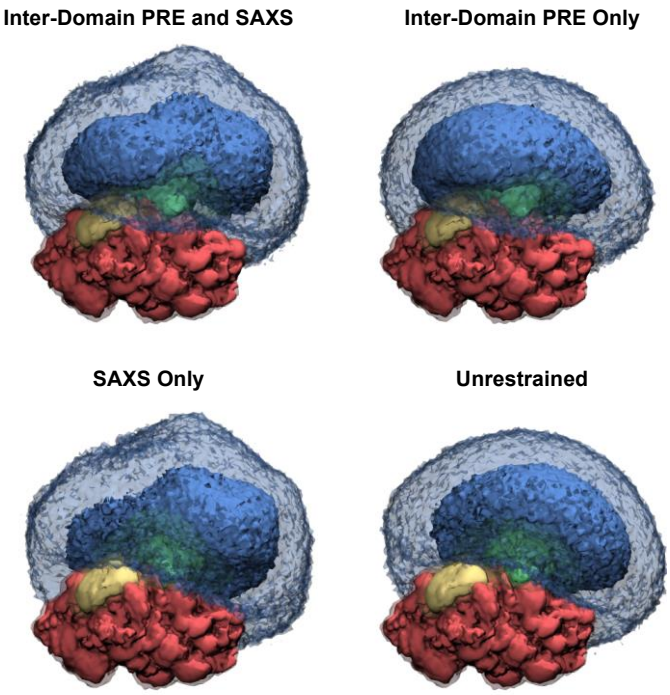

Figure 4.34. Probability distributions of inter-domain orientations in ensembles calculated using different restraint sets. The CVNH and LysM domains are shown in red and blue, respectively, the inter-domain linker in green, and MTSL paramagnetic tag in yellow. Structures were best fit to the CVNH domain coordinates, and the simulation cell was divided into 1-$Å^3$ bins. Solid contours represent bins occupied by a heavy atom in at least 1% of the ensemble, while transparent contours represent bins occupied in at least 0.1% of the ensemble.

5.0 CONCLUSIONS AND FUTURE DIRECTIONS

As discussed in Chapter 1, multi-domain proteins are challenging to characterize with traditional experimental structural biology techniques, but are amenable to study using integrative methods that combine the results of experiments with computer simulations. However, these methods are subject to the accuracy of the force fields used for the simulations, and well-characterized benchmark systems against which to validate models are therefore needed. The flexibly linked two-domain protein MoCVNH3 is an excellent system for such validation. Prior to the work described here, our group had solved the structure of MoCVNH3 using solution NMR, finding that while the two domains have well-defined structures, there is no fixed inter-domain orientation between them. This prevented crystallization, although a variant with shortened inter-domain linkers was successfully crystallized and its structure solved. The conformational sampling of flexible systems such as MoCVNH3 in computer simulations depends on the balance of protein-protein and protein-solvent interactions in the force fields used. MoCVNH3 can be used to simultaneously validate force fields' accuracy in (i) maintaining the structures of folded proteins and (ii) balancing protein-protein and protein-solvent interactions. After running an MD simulation of MoCVNH3 using the recently-developed AMBER ff99SB-ILDN force field and TIP4P-Ew water model, I found that the two domains stuck together in a single inter-domain orientation, in contrast to expectations based on our solution NMR data. Examination of the contacts between the two domains revealed a series of salt bridges, leading me to suspect that the force field was overstabilizing these interactions.

I therefore investigated the strength of salt bridge interactions in biomolecular force fields, as described in Chapter 2. Using minimal model systems, consisting of side-chain analogues of

three oppositely charged amino acid pairs, I compared an assortment of current biomolecular force fields and water models both to each other and to experiment. I found considerable variation between the force fields, and that salt bridge interactions were overstabilized by all of them. I confirmed that our results extend to the complete amino acids by also simulating blocked arginine and aspartate dipeptides. With this system, I found that the AMBER ff13α force field offered apparent improvement. Importantly, this force field was developed using a new approach to fitting nonbonded parameters, the Implicitly Polarized Charge (IPolQ) method, that offered a promising path towards more accurate force fields without resorting to more expensive function forms (Figure 5.1).

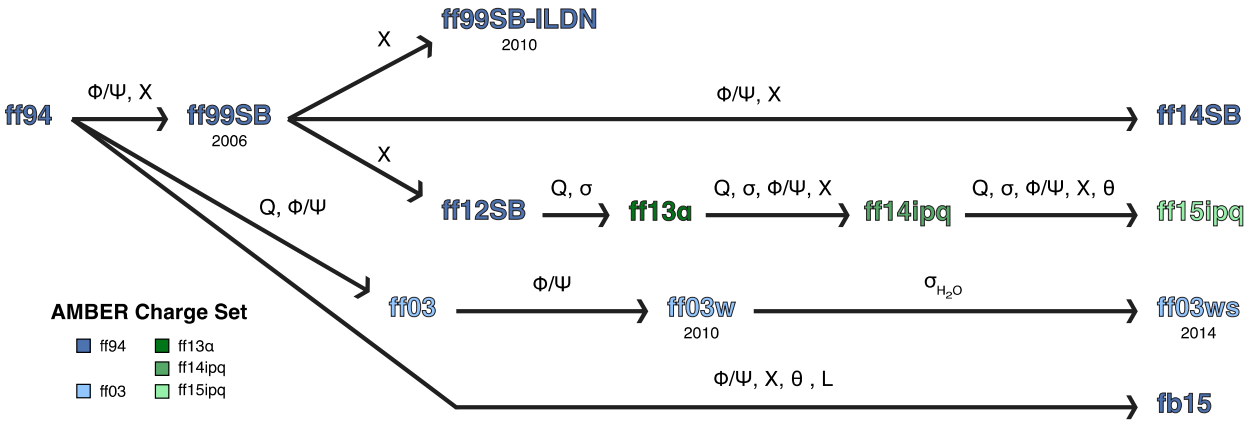

Figure 5.1. Phylogeny of selected AMBER force fields mentioned in this text. At each step changes were made to a subset of parameters including charges (Q), van der Waals parameters (σ), backbone torsions (Φ/Ψ), side-chain torsions (X), angles (θ), and bonds (L). Force field names are colored based on their charge set; the three IPolQ charge sets (green) were derived using the same philosophy but have minor differences. For force fields whose name does not reflect their approximate year of release, the year of publication is listed below their name.

As AMBER ff13α was extended to a complete force field, ff14ipq, changes were made that we discovered had led to drastic overstabilization of salt bridge interactions. I collaborated with ff14ipq's designers to develop an improved version, AMBER ff15ipq, that yields accurate salt

bridge interactions relative to other fixed-charge force fields, as described in Chapter 3. However, I found that ff15ipq still underperforms relative to more expensive force fields that include explicit polarization. Alongside ff15ipq's updated nonbonded parameters, we improved the bonded parameters by quadrupling the number of quantum mechanical target conformations and decoupling the parameters of different amino acids. Through extensive simulations of peptides and proteins we found that ff15ipq yields extremely encouraging reproduction of NMR observables. In particular, ff15ipq reproduces the J-coupling constants of the Ala_5 peptide more accurately than contemporary force fields that have been fit specifically to this empirical result, providing a major validation of the IPolQ method of parameterization. In addition, when paired with the SPC/E_b water model with which the force field was developed, ff15ipq yields accurate ^{15}N R_1 and R_2 relaxation parameters. These parameters report on the kinetics of each residue, and when paired with the thermodynamics provided by the conformational sampling provide a composite view of the per-residue accuracy of the force field. We demonstrated that ff15ipq accurately retains the folded structures of globular proteins over multiple microseconds, and captures the folding of disordered peptides upon interaction with their binding partners.

After developing an updated force field that addressed the suspected artifacts observed in my initial simulation, I simulated MoCVNH3 and a series of mutants containing reduced length inter-domain linkers using AMBER ff15ipq/SPC/E_b and another force field/water model combination, CHARMM22*/TIP4P-D. Alongside these simulations we collected additional experimental data, such as SAXS for validating the overall structure of the two-domain system. I found that despite the adjustments made to the salt bridges, my ff15ipq/SPC/E_b simulation remained stuck in a single inter-domain orientation for most of the simulation. However, my CHARMM22*/TIP4P-D simulation did not, due to adjustments to the balance of protein-protein

and protein-water interactions made in the TIP4P-D water model. Thus, while our adjustments to the strength of salt bridge interactions may be valid and necessary, the adjustments were not sufficient to completely fix the underlying deficiency as changes to the water model were also necessary. My ff15ipq/SPC/E_b simulations did, however, retain the internal structure of the CVNH and LysM domains more reliably than our CHARMM22*/TIP4P-D simulations, reinforcing the promise of the IPolQ parameterization method.

I established with this work that MoCVNH3 was a valuable model system, and collected further experimental data in order to determine the overall structure of the two-domain system. In particular, I collected PRE data to experimentally characterize the distribution of inter-domain orientations, attaching a paramagnetic tag to the CVNH domain and measuring the effect on resonances in the LysM domain. Our PRE data revealed a pair of mutually exclusive contact sites on opposite sides of the LysM domain; both sites cannot approach the paramagnetic tag on the CVNH domain simultaneously, demonstrating that the inter-domain orientations must exchange rapidly. Using our experimental SAXS and PRE data for MoCVNH3, we calculated a structural ensemble of inter-domain orientations. We found that while the two domains of MoCVNH3 indeed have no fixed inter-domain orientation, it is still possible to resolve differences in population among the possible inter-domain orientations.

The results of the work carried out during my thesis research have already led to tangible improvements in biomolecular force fields. In our comparison of salt bridge interactions between different force fields, we found that the CHARMM27 force field was one of the more strongly overstabilizing. The latest iteration in this force field lineage, CHARMM36m,[204] has been revised in light of our results to have weaker salt bridge interactions. Similar adjustments have been made to variants of the AMBER ff99SB-ILDN force field, with which we first observed the issue.[94,95]

However, the post hoc corrections made to these force fields, the bulk of whose nonbonded parameters have not changed since the 1990s, leave AMBER ff15ipq as the only force field to have all its nonbonded and bonded parameters fit consistent with its correction to the strength of salt bridges.

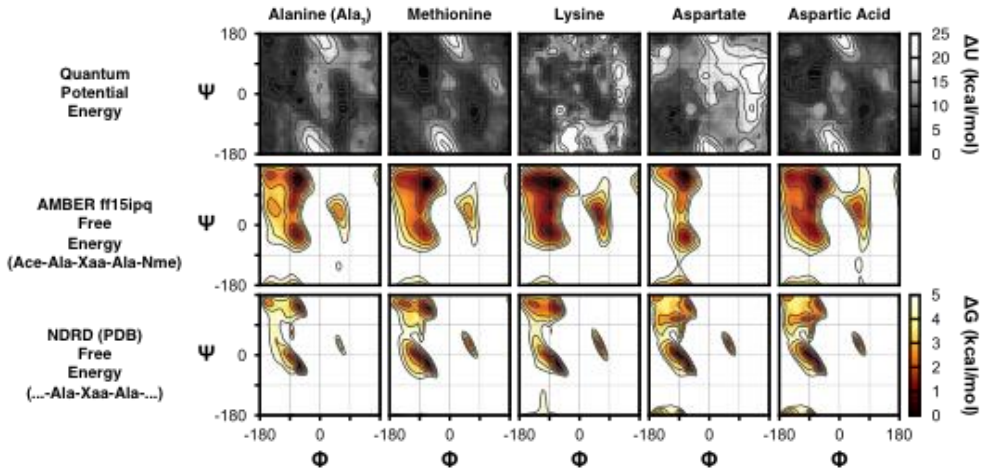

Figure 5.2. Comparison of QM target, AMBER ff15iq model, and empirical energy surfaces for Ala, Met, Lys, Asp, and Ash (aspartic acid). Top, target QM potential energy surfaces (PES) for the Ace-Ala_3-Nme tetrapeptide and the Ace-Xaa-Nme depeptides of Met, Lys, Asp, and Ash. Middle, AMBER ff15ipq free energy surface (FES) for the Ace-Ala-Xaa-Ala-Nme tetrapeptides of each amino acid. Bottom, empirical FES of …-Ala-Xaa-Ala-… for each amino acid, derived from statistical analysis of the loop regions of high-resolution crystal structures.[143] The Ala_3 tetrapeptide QM potential energy surface (PES) contains features not present in the Met dipeptide PES (serving as a representative example of all neutral dipeptide PES), and these features are expected to be present based on the empirical FES. The QM PES of the charged dipeptides of Lys and Asp bear little resemblance to the empirical FES.

The main direction for future development of the AMBER IPolQ lineage of force fields will be the addition of virtual sites. These will enable more accurate reproduction of the distribution of charge within molecules without altering the inexpensive functional form of the force field. However, analysis of our results carried out since the work described in Chapter 3 was completed has revealed several potential directions for improvement that do not require virtual sites. Detailed

analysis of the target QM potential energy surface (PES) shows that the PES for tetrapeptides contain many more of the expected features than do the PES for dipeptides, which comprised most of the fitting set (Figure 5.2). This is likely why ff15ipq's improvements in Ala are so stark, as manifested in its excellent Ala_5 J-couplings, while the improvements for other residues are more ambiguous. Applying the IPolQ method to a fitting set that included exclusively tetrapeptides could therefore yield considerable improvement. Also clear from examination of the QM target PES is that the PES of charged residues have few of the expected features based on the empirical conformations of these residues. This is likely why, in ff15ipq, negatively charged residues Asp and Glu exhibit much more restricted wells, while positively charged residues Arg and Lys exhibit little discrimination between conformations. These observations extend to our simulations of benchmark peptides and proteins: nearly all deviations from native structure involved charged residues. It appears that the QM PES of charged amino acids may simply be a poor target against which to fit force fields for the solution phase. A key future direction will therefore be to explore adjustments to the IPolQ protocol to enable the omission of QM calculations on charged systems from the fitting set.

As force fields continue to improve, our results described in Chapter 4 for MoCVNH3 will provide a valuable reference data set against which to validate future force fields, including polarizable force fields. While polarizable force fields exceed nonpolarizable force fields for selected metrics such as salt bridge stability, polarizable force fields have other limitations that will need to be addressed before systems like MoCVNH3 will be useful enough to justify the cost of simulating them. For example, in tests of polarizable force fields, the loss of native structure of GB3 and ubiquitin observed within tens of nanoseconds[108] exceeds that which occurs after multiple microseconds for nonpolarizable force fields. Once such issues are addressed in single-domain

protein systems, MoCVNH3 will become a valuable validation system for polarizable force fields. Overall, our work on MoCVNH3 has laid important groundwork for the characterization of the structure and dynamics other flexibly linked multi-domain proteins using integrative structural biology methods.

6.0 APPENDIX

6.1 ADDITIONAL PUBLICATIONS

In addition to the three first-author manuscripts reproduced above, over the course of this work smaller contributions were made to two additional publications. The abstracts of each are reproduced below.

6.1.1 Characterization of the Mo-0v reduced linker-length variant of MoCVNH3

Published as: Koharudin, L. M. I., Debiec, K. T.; Gronenborn, A. M. Structural insight into fungal cell wall recognition by a CVNH protein with a single LysM Domain. *Structure*, 2015, 23, 2143-2154.

MGG_03307 is a lectin isolated from *Magnaporte oryzae*, a fungus that causes devastating rice blast disease. Its function is associated with protecting *M. oryzae* from the host immune response in plants. To provide the structural basis of how MGG_03307 protects the fungus, crystal structures of its CVNH- LysM module were determined in the absence and presence of GlcNAc-containing cell wall chitin constituents, which can act as pathogen-associated molecular patterns. Our structures revealed that glycan binding is accompanied by a notable conformational change in the LysM domain and that $GlcNAc_3$ and $GlcNAc_4$ are accommodated similarly. $GlcNAc_5$ and $GlcNAc_6$ interact with the LysM domain in multiple conformations, as evidenced by solution nuclear magnetic resonance studies. No dimerization of MoCVNH3 via its LysM domain was

observed upon binding to $GlcNAc_6$, unlike in multiple LysM domain-containing proteins. Importantly, we define a specific consensus binding mode for the recognition of GlcNAc oligomers by single LysM domains.

6.1.2 Validation of the IPolQ method of force field parameterization

Submitted to *Journal of Chemical Physics* for publication as: Cerutti, D. S.; Debiec, K. T.; Case, D. A.; Chong L. T. Significance of the charge model in biomolecular force field design.

The ff15ipq protein force field is a fixed-charge model built by automated tools based on the two charge sets of the Implicitly Polarized Charge method: one set for deriving bonded parameters and the other for running simulations. The duality is intended to treat water-induced electronic polarization with an understanding that fitting data for bonded parameters will come from quantum mechanical calculations in the gas phase. In this study, we compare ff15ipq to two alternatives produced with the same fitting software and a large subset of the same data, but following more conventional methods for tailoring bonded parameters (harmonic angle terms and torsion potentials) to the charge model. The first, f15ipq-Q_{solv}, derives bonded parameters in the context of the ff15ipq solution-phase charge set, and ff15ipq-Q_{vac}, which takes ff15ipq's bonded parameters and runs simulations with ff15ipq's vacuum-phase charge set used to derive those parameters. The IPolQ charge model and associated protocol for deriving bonded parameters are shown to be an incremental improvement over protocols that do not account for the material phases of each source of their fitting data. Both force fields incorporating the polarized charge set depict stable globular proteins and have varying degrees of success modeling the metastability of short (5 to 19 residue) peptides. In this particular case, ff15ipq-Q_{solv} increases stability in a number of

α-helices, correctly obtaining 70% helical character in the K19 system at 275K and showing appropriately diminishing content up to 325K, but overestimating the helical fraction of AAQAA3 by 50% or more, forming long-lived a-helices in simulations of a β-hairpin, and increasing the likelihood that the disordered p53 N-terminal peptide will also form a helix. In contrast, ff15ipq-Q_{Vac} incorrectly depicts globular protein unfolding in numerous systems tested, including Trp cage, villin, lysozyme, and GB3, and does not perform any better than ff15ipq or ff15ipq-Qs_{olv} in tests on short peptides. We analyze the free energy surfaces of individual amino acid dipeptides and the electrostatic potential energy surfaces of each charge model to explain the differences.

6.2 SOFTWARE DEVELOPED

6.2.1 MolDynPlot

MolDynPlot is a Python package for analyzing and plotting data from MD simulations and biophysical experiments. MolDynPlot reads in data from sources including AmberTools' cpptraj program for MD simulations,[138] the PRIMUS, CRYSOL, and FOXS programs for SAXS data,[188,190,205] and the Nmrglue library for NMR data.[206] Once read in from their original formats, data are organized into a consistent data structure provided by the pandas library,[207] which may be written to and read from either text or the efficient HDF5 database format.[208] MolDynPlot integrates data processing and analysis, supporting operations including the calculation of averages (with block-average standard errors)[64] and probability distributions over the course of time series. MolDynPlot includes tools for plotting time series, probability distributions, SAXS curves, NMR spectra, and several other formats using the Matplotlib package,[209] which is extended through the

addition of a framework for configuring plots using the YaML text format.[210] Plot configuration is specified using a hierarchical structure that makes precise per-figure, per-plot, and per-dataset configuration possible without needing to write additional Python code. MolDynPlot additionally supports writing amino acid sequence-based datasets onto the β or occupancy column of PDB files, which may be subsequently used by the Visual Molecular Dynamics program to draw protein structures whose coloration corresponds to the data.[173] MolDynPlot is freely available on GitHub under a 3-clause BSD license.[211]

6.2.2 Ramaplot

Ramaplot is a Python package for the generation of Ramachandran plots, used for visualizing probability, free energy, potential energy, or other quantities as a function of the backbone Φ/Ψ angles of an amino acid. Like MolDynPlot, Ramaplot is equipped to parse, analyze, and plot output from AmberTools' cpptraj program, which includes complementary functions for analyzing MD simulations.[138] Ramaplot is also equipped to work with data from the Weighted Histogram Analysis Method (WHAM),[119] an enhanced sampling technique useful for rapidly quantifying the backbone Φ/Ψ conformational preferences of amino acids. For validating the results of MD simulations, Ramaplot supports experimental data from two datasets generated through statistical analysis of known protein structures. The Neighbor-Dependent Ramachandran Distribution (NDRD) dataset includes Φ/Ψ distributions for each amino acid derived from the loop regions of high-resolution structures, including consideration of the influence of adjacent residues.[143] The Conformation-Dependent Library (CDL) dataset includes the average values of the backbone heavy atom bond lengths, angles, and ω torsion as a function of backbone Φ/Ψ.[212,213] Ramaplot supports analysis functions including the subtraction of distributions and the calculation of the

populations of different states within Φ/Ψ space. Ramaplot is freely available on GitHub under a 3-clause BSD license.[214]

BIBLIOGRAPHY

(1) Levitt, M. Nature of the Protein Universe. *Proc. Natl. Acad. Sci.* **2009**, *106* (27), 11079–11084.

(2) Ekman, D.; Björklund, Å.˚K.; Frey-Skött, J.; Elofsson, A. Multi-Domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions. *J. Mol. Biol.* **2005**, *348* (1), 231–243.

(3) Bashton, M.; Chothia, C. The Generation of New Protein Functions by the Combination of Domains. *Structure* **2007**, *15* (1), 85–99.

(4) Bhaskara, R. M.; de Brevern, A. G.; Srinivasan, N. Understanding the Role of Domain-Domain Linkers in the Spatial Orientation of Domains in Multi-Domain Proteins. *J. Biomol. Struct. Dyn.* **2013**, *31* (12), 1467–1480.

(5) Papaleo, E.; Saladino, G.; Lambrughi, M.; Lindorff-Larsen, K.; Gervasio, F. L.; Nussinov, R. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* **2016**, *116* (11), 6391–6423.

(6) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. Conformational Selection or Induced Fit: A Flux Description of Reaction Mechanism. *Proc. Natl. Acad. Sci.* **2009**, *106* (33), 13737–13741.

(7) Aroul-Selvam, R.; Hubbard, T.; Sasidharan, R. Domain Insertions in Protein Structures. *J. Mol. Biol.* **2004**, *338* (4), 633–641.

(8) van den Bedem, H.; Fraser, J. S. Integrative, Dynamic Structural Biology at Atomic Resolution—it's about Time. *Nat. Methods* **2015**, *12* (4), 307–318.

(9) Bernadó, P. Effect of Interdomain Dynamics on the Structure Determination of Modular Proteins by Small-Angle Scattering. *Eur. Biophys. J.* **2010**, *39* (5), 769–780.

(10) Madl, T.; Gabel, F.; Sattler, M. NMR and Small-Angle Scattering-Based Structural Analysis of Protein Complexes in Solution. *J. Struct. Biol.* **2011**, *173* (3), 472–482.

(11) Perkins, S. J.; Wright, D. W.; Zhang, H.; Brookes, E. H.; Chen, J.; Irving, T. C.; Krueger, S.; Barlow, D. J.; Edler, K. J.; Scott, D. J.; Terrill, N. J.; King, S. M.; Butler, P. D.; Curtis, J. E.; A., B. N.; D., S.; S., J.; J., H. M.; A., M. J.; B., B. R.; X., Z.; J., S.; M., L. P. E.; J., M.; M., F.; Jr, M. A. D.; K., B. M.; M., W. J.; A., K. M.; J., P. S.; H., B. E.; N., A.; E., C. J.; S., M.; R., S.; M., P.; A., C. C.; A., C.; M., C. C.; E., C. J.; S., K.; D., F.; A., C. C.; K., D. E.;

O., W.; A., C.; A., N. M.; E., C. J.; R., R. M.; S., K.; A., C. T.; D., F.; J., C.; L., B. C.; J., K.; J., C.; W., I.; L., B. C.; C., C. P.; S., H. J.; J., C. N.; H., Z.; S., K.; J., L. H.; R., K. R.; B., K.; R., K. S.; J., T. M.; A., M.; E., C. J.; E., C. J.; S., R.; H., N.; S., K.; K., D. S. A.; E., C. J.; W., R.; K., C. P.; M., C. R.; J., L.; S., K.; A., R.; D., D. C.; I., G.; D., H.; M., B.; M., J.; C., A.; A., P.-M.; G., M.; R., C.; J., D. T.; P., C.; H., L.; E., N. J.; H., J. J.; G., K.; A., B. N.; J., D. T.; E., N. J.; A., M. J.; A., B. N.; G., E.; F., M.; F., B.; C., S.; J., P.; P., F.; J., A.; A., C.; M., G.; L., H.; E., B.; P., S.; J., S. D.; M., H. R.; J., P.; Z., R.; T., N.; L., S. J.; K., H. R.; E., R. S.; D., T.; E., T. A.; J., T.; M., K. S.; B., H. A.; R., B. E.; J., B. P.; K., H. G.; W., W. D.; L., V. O.; E., R. L.; M., P.; C., Y. S.; J., G.; K., M.; J., B.; J., P. S.; W., H.; A., D.; K., S.; L., H. G.; L., M. A.; M., H.; P., R. R.; L., P. F.; E., T. S.; Jr, J. F. E.; S., C.; A., F. K.; C., H. R.; S., Y.; W., S. J.; D., D. B.; W., A. M. W.; A., T. J.; C., I. E.; P., F. M.; B., J.-G.; C., P.; I., S. D.; P., B.; J., F.-R.; S., J.; T., K.; G., I. V.; W., I.; S., K.; J., G.; B., M.; J., P. S.; J., K. C.; S., H. J.; J., K.; G., H.; J., L.; S., D.; R., L.; L., R. S.; P., S. M. S.; J., M. A.; D., M. A.; O., M. M.; J., C. W.; D., B.; A., S. D. B.; J., P. S.; D., P.; F., M.; C., P. M.; L., A.; K., M.; M., P.; L., H. G.; M., H.; Y., P.; E., C. J.; X., F.; A., W. S.; J., P. S.; J., P. S.; R., N.; K., L.; S., K.; Y., A.; J., P. S.; S., N. A.; J., S. B.; A., F.; J., P. S.; I., O. A.; N., F. A.; A., B.; E., G. H.; B., F. P.; J., P. S.; I., O. A.; R., N.; K., L.; A., B.; P., P.; V., P. M.; D., F.; V., S. A.; G., T.; G., K. A.; M., G.; C., G.; T., M. H. D.; V., K. P.; I., S. D.; F., P. E.; D., G. T.; C., H. C.; S., C. G.; M., G. D.; C., M. E.; E., F. T.; C., P. J.; R., B.; W., W.; J., G.; E., T.; E., V.; C., C.; D., S. R.; L., K.; K., S.; F., P.; H., O.; S., D.; P., K.; M., D.; D., P. C.; M., H.; L., H. G.; A., T. J.; P., R. R.; A., T. J.; E., R. L.; K., H. G.; J., G.; K., H. R.; A., D. P.; J., P. S.; E., R. L.; K., H. G.; J., G.; K., H. R.; A., D. P.; J., P. S.; A., R.; F., F.; L., F.; A., G.; J., H.; C., V.; E., B. C.; P., P.; S., M.; M., R.; I., S. D.; F., C.; B., R.; C., K. Y.; G., H.; A., Š.; L., B. T.; D., S.-D.; M., H.; A., S.; D., S.-D.; M., H.; A., T. J.; A., S.; F., S.; M., B.; D., S.; C., B.; J., K. M. H.; W., W.; L., W.; J., W.; Q., G.; Y., S.; C., W. M.; E., C. J.; E., W. A.; S., C.; J., T.; D., W. M.; C., W.; T., B.; A., P.; S., S.; M., T.; A., R.; M., W.; W., W. D.; J., P. S.; S., Y.; L., B.; L., M.; B., R.; S., Y.; S., P.; L., M.; B., R.; Z., Z.; Y., C. J.-M.; S., A.; B., W.; K., H.; S., S. G.; C., D. Atomistic Modelling of Scattering Data in the Collaborative Computational Project for Small Angle Scattering (CCP-SAS). *J. Appl. Crystallogr.* **2016**, *49* (6), 1861–1875.

(12) Zhao, D.; Wang, X.; Peng, J.; Wang, C.; Li, F.; Sun, Q.; Zhang, Y.; Zhang, J.; Cai, G.; Zuo, X.; Wu, J.; Shi, Y.; Zhang, Z.; Gong, Q. Structural Investigation of the Interaction between the Tandem SH3 Domains of c-Cbl-Associated Protein and Vinculin. *J. Struct. Biol.* **2014**, *187* (2), 194–205.

(13) Kikhney, A. G.; Svergun, D. I. A Practical Guide to Small Angle X-Ray Scattering (SAXS) of Flexible and Intrinsically Disordered Proteins. *FEBS Lett.* **2015**, *589* (19), 2570–2577.

(14) Schneidman-duhovny, D.; Rossi, A.; Avila-sakar, A.; Kim, S. J.; Strop, P.; Liang, H.; Krukenberg, K. A.; Liao, M.; Kim, H. M.; Sobhanifar, S.; Rajpal, A.; Pons, J.; David, A.;

Cheng, Y.; Sali, A. A Method for Integrative Structure Determination of Protein-Protein Complexes. *Bioinformatics* **2012**, *28* (24), 3282–3289.

(15) Tainer, J. A. X-Ray Scattering (SAXS) Combined with Crystallography and Computation: Defining Accurate Macromolecular Structures, Conformations and Assemblies in Solution. *Biophys. J.* **2011**, *100* (3), 38a.

(16) Koharudin, L. M. I.; Furey, W.; Gronenborn, A. M. Novel Fold and Carbohydrate Specificity of the Potent Anti-HIV Cyanobacterial Lectin from Oscillatoria Agardhii. *J. Biol. Chem.* **2011**, *286* (2), 1588–1597.

(17) Koharudin, L. M. I.; Debiec, K. T.; Gronenborn, A. M. Structural Insight into Fungal Cell Wall Recognition by a CVNH Protein with a Single LysM Domain. *Structure* **2015**, *23* (11), 2143–2154.

(18) Koharudin, L. M. I.; Gronenborn, A. M. Structural Basis of the Anti-HIV Activity of the Cyanobacterial Oscillatoria Agardhii Agglutinin. *Structure* **2011**, *19* (8), 1170–1181.

(19) Martin-Urdiroz, M.; Oses-Ruiz, M.; Ryder, L. S.; Talbot, N. J. Investigating the Biology of Plant Infection by the Rice Blast Fungus Magnaporthe Oryzae. *Fungal Genet. Biol.* **2016**, *90* (March 2015), 61–68.

(20) Percudani, R.; Montanini, B.; Ottonello, S. The Anti-HIV Cyanovirin-N Domain Is Evolutionarily Conserved and Occurs as a Protein Module in Eukaryotes. *Proteins* **2005**, *60* (4), 670–678.

(21) de Jonge, R.; Thomma, B. P. H. J. Fungal LysM Effectors: Extinguishers of Host Immunity? *Trends Microbiol.* **2009**, *17* (4), 151–157.

(22) Koharudin, L. M. I.; Viscomi, A. R.; Montanini, B.; Kershaw, M. J.; Talbot, N. J.; Ottonello, S.; Gronenborn, A. M. Structure-Function Analysis of a CVNH-LysM Lectin Expressed during Plant Infection by the Rice Blast Fungus Magnaporthe Oryzae. *Structure* **2011**, *19* (5), 662–674.

(23) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science (80-.).* **1998**, *282*, 740–744.

(24) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana-Agostinetti, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science (80-.).* **2010**, *330* (October), 341–346.

(25) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, H. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.;

McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **2008**, *51* (7), 91–97.

(26) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A. E.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725.

(27) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665–9678.

(28) Lindorff-Larsen, K.; Piana-Agostinetti, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950–1958.

(29) Wong, V.; Case, D. A. Evaluating Rotational Diffusion from Protein MD Simulations. *J. Phys. Chem. B* **2008**, *112*, 6013–6024.

(30) Donald, J. E.; Kulp, D. W.; DeGrado, W. F. Salt Bridges: Geometrically Specific, Designable Interactions. *Proteins* **2011**, *79* (3), 898–915.

(31) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; XIong, G.; Zhang, W.; Yang, R.; Piotr, C.; Luo, R.; Lee, T.; Caldwell, J. W.; Wang, J.; Kollman, P. A. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999–2012.

(32) Piana-Agostinetti, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–L49.

(33) Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *J. Phys. Chem. B* **2013**, *117*, 2328–2338.

(34) Novotny, J.; Sharp, K. A. Electrostatic Fields in Antibodies and Antibody/antigen Complexes. *Prog. Biophys. Mol. Biol.* **1992**, *58*, 203–224.

(35) Hendsch, Z. S.; Tidor, B. Do Salt Bridges Stabilize Proteins? A Continuum Electrostatic Analysis. *Protein Sci.* **1994**, *3* (2), 211–226.

(36) Elcock, A. H. The Stability of Salt Bridges at High Temperatures: Implications for Hyperthermophilic Proteins. *J. Mol. Biol.* **1998**, *284* (2), 489–502.

(37) Hendsch, Z. S.; Tidor, B. Electrostatic Interactions in the GCN4 Leucine Zipper: Substantial Contributions Arise from Intramolecular Interactions Enhanced on Binding. *Protein Sci.*

1999, *8* (7), 1381–1392.

(38) Sheinerman, F. B.; Honig, B. On the Role of Electrostatic Interactions in the Design of Protein-Protein Interfaces. *J. Mol. Biol.* **2002**, *318* (1), 161–177.

(39) Salari, R.; Chong, L. T. Desolvation Costs of Salt Bridges across Protein Binding Interfaces: Similarities and Differences between Implicit and Explicit Solvent Models. *J. Phys. Chem. Lett.* **2010**, *1*, 2844–2848.

(40) Salari, R.; Chong, L. T. Effects of High Temperature on Desolvation Costs of Salt Bridges across Protein Binding Interfaces: Similarities and Differences between Implicit and Explicit Solvent Models. *J. Phys. Chem. B* **2012**, *116*, 2561–2567.

(41) Saigal, S.; Pranata, J. Monte Carlo Simulations of Guanidinium Acetate and Methylammonium Acetate Ion Pairs in Water. *Bioorg. Chem.* **1997**, *25* (1), 11–21.

(42) Rozanska, X.; Chipot, C. Modeling Ion-Ion Interactions in Proteins: A Molecular Dynamics Free Energy Calculation of the Guanidinum-Acetate Assciation. *J. Chem. Phys.* **2000**, *112* (22), 9664–9691.

(43) Masunov, A.; Lazaridis, T. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. *J. Am. Chem. Soc.* **2003**, *125* (15), 1722–1730.

(44) Henin, J.; Chipot, C. Overcoming Free Energy Barriers Using Unconstrained Molecular Dynamics Simulations. *J. Chem. Phys.* **2004**, *121* (7), 2904–2914.

(45) Thomas, A. S.; Elcock, A. H. Molecular Simulations Suggest Protein Salt Bridges Are Uniquely Suited to Life at High Temperatures. *J. Am. Chem. Soc.* **2004**, *126* (7), 2208–2214.

(46) Yu, Z.; Jacobson, M. P.; Josovitz, J.; Rapp, C. S.; Friesner, R. A. First-Shell Solvation of Ion Pairs: Correction of Systematic Errors in Implicit Solvent Models. *J. Phys. Chem. B* **2004**, *108* (21), 6643–6654.

(47) Tan, C.; Yang, L.; Luo, R. How Well Does Poisson - Boltzmann Implicit Solvent Agree with Explicit Solvent ? A Quantitative Analysis. **2006**, 18680–18687.

(48) Thomas, A. S.; Elcock, A. H. Direct Observation of Salt Effects on Molecular Interactions through Explicit-Solvent Molecular Dynamics Simulations: Differential Effects on Electrostatic and Hydrophobic Interactions and Comparisons to Poisson-Boltzmann Theory. *J. Am. Chem. Soc.* **2006**, *128* (24), 7796–7806.

(49) Zhu, S.; Elcock, A. H. A Complete Thermodynamic Characterization of Electrostatic and Hydrophobic Associations in the Temperature Range 0 to 100 °C from Explicit-Solvent Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2010**, *6* (4), 1293–1306.

(50) Andrews, C. T.; Elcock, A. H. Molecular Dynamics Simulations of Highly Crowded Amino Acid Solutions: Comparisons of Eight Different Force Field Combinations with Experiment and with Each Other. *J. Chem. Theory Comput.* **2013**.

(51) Springs, B.; Haake, P. Equilibrium Constants for Association of Guanidinium and Ammonium Ions with Oxyanions. *Bioorg. Chem.* **1977**, *6* (2), 181–190.

(52) Martinez, L.; Andrade, R.; Birgin, E. G.; Martinez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30* (13), 2157–2164.

(53) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(54) Roux, B. Valence Selectivity of the Gramicidin Channel: A Molecular Dynamics Free Energy Perturbation Study. *Biophys. J.* **1996**, *71*, 3177–3185.

(55) Chandrasekhar, J.; Spellmeyer, D. C.; Jorgensen, W. L. Energy Component Analysis for Dilute Aqueous Solutions of Li+, Na+, F-, and Cl- Ions. *J. Am. Chem. Soc.* **1984**, *106* (13), 903–910.

(56) Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. Evaluating the Strength of Salt Bridges: A Comparison of Current Biomolecular Force Fields. *J. Phys. Chem. B* **2014**, *118*, 6561–6569.

(57) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossváry, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the 2006 ACM/IEEE SC|06 Conference*; 2006; pp 1–13.

(58) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press, 1989.

(59) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189.

(60) Krautler, V.; van Gunsteren, W. F.; Hunenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508.

(61) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(62) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. Nose–Hoover Chains: The Canonical Ensemble via Continuous Dynamics. *J. Chem. Phys.* **1992**, *97* (4), 2635–2643.

(63) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian Split Ewald:

A Fast Ewald Mesh Method for Molecular Simulation. *J. Chem. Phys.* **2005**, *122* (5), 1–13.

(64) Flyvbjerg, H.; Petersen, H. G. Error Estimates on Averages of Correlated Data. *J. Chem. Phys.* **1989**, *91* (1), 461–466.

(65) Neumann, M.; Steinhauser, O. On the Calculation of the Frequency-Dependent Dielectric Constant in Computer Simulations. *Chem. Phys. Lett.* **1983**, *102* (6), 508–513.

(66) Yang, L.; Weerasinghe, S.; Smith, P. E.; Pettitt, B. M. Dielectric Response of Triplex DNA in Ionic Solution from Simulations. *Biophys. J.* **1995**, *69* (4), 1519–1527.

(67) MacKerell Jr., A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415.

(68) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26* (16), 1752–1780.

(69) Tanford, C. The Association of Acetate with Ammonium and Guanidinium Ions. *J. Am. Chem. Soc.* **1954**, *76*, 945–946.

(70) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(71) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902–1921.

(72) Abascal, J. L. F.; Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123* (23), 234505.

(73) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(74) Fernández, D. P.; Goodwin, A. R. H.; Lemmon, E. W.; Levelt Sengers, J. M. H.; Williams, R. C. A Formulation for the Static Permittivity of Water and Steam at Temperatures from 238 K to 873 K at Pressures up to 1200 MPa, Including Derivatives and Debye–Hückel Coefficients. *J. Phys. Chem. Ref. Data* **1997**, *26* (4), 1125.

(75) Grossfield, A.; Ren, P.; Ponder, J. W. Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671–15682.

(76) Whitfield, T. W.; Varma, S.; Harder, E.; Lamoureux, G.; Rempe, S. B.; Roux, B.

Theoretical Study of Aqueous Solvation of K+ Comparing Ab Initio, Polarizable, and Fixed-Charge Models. *J. Chem. Theory Comput.* **2007**, *3*, 2068–2082.

(77) Liang, T.; Walsh, T. R. Molecular Dynamics Simulations of Peptide Carboxylate Hydration. *Phys. Chem. Chem. Phys.* **2006**, *8* (38), 4410–4419.

(78) Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-Accelerated Molecular Modeling Coming of Age. *J. Mol. Graph. Model.* **2010**, *29*, 116–125.

(79) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without Compromise - A Mixed Precision Model for GPU Accelerated Molecular Dynamics Simulations. *Comput. Phys. Commun.* **2013**, *184* (2), 374–380.

(80) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(81) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(82) Li, D.-W.; Brüschweiler, R. Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. *J. Chem. Theory Comput.* **2011**, *7* (6), 1773–1782.

(83) Li, D.-W.; Brüschweiler, R. NMR-Based Protein Potentials. *Angew. Chemie* **2010**, *49* (38), 6778–6780.

(84) Jiang, F.; Zhou, C. Y.; Wu, Y. Residue-Specific Force Field Based on the Protein Coil Library . RSFF2: Modification of AMBER ff99SB. *J. Phys. Chem. B* **2014**, *119*, 1035–1047.

(85) Best, R. B.; Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* **2010**, *114* (46), 14916–14923.

(86) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.

(87) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell Jr., A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone Φ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(88) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015**, *11*, 3499–3509.

(89) Jiang, F.; Zhou, C. Y.; Wu, Y. D. Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L. *J. Phys. Chem. B* **2014**, *118* (25), 6983–6998.

(90) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett.* **2014**, *5* (11), 1885–1891.

(91) Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 4515–4534.

(92) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J. Phys. Chem. B* **2017**, acs.jpcb.7b02320.

(93) Case, D. A.; Berryman, J. T.; Betz, R. M.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Götz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P. A.; Kaus, J. W.; Kovalenko, A.; Lee, T.; Le Grand, S.; Li, P.; Luchko, T.; Luo, R.; Madej, B. D.; Merz, K. M.; Monard, G.; Needham, H.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A. E.; Salomon-Ferrer, R.; Simmerling, C.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; York, D. M.; Kollman, P. A. AMBER 2015. University of California, San Francisco 2015.

(94) Xue, Y.; Yuwen, T.; Zhu, F.; Skrynnikov, N. R. Role of Electrostatic Interactions in Binding of Peptides and Intrinsically Disordered Proteins to Their Folded Targets. 1. NMR and MD Characterization of the Complex between the c-Crk N-SH3 Domain and the Peptide Sos. *Biochemistry* **2014**, *53* (41), 6473–6495.

(95) Xue, Y.; Yuwen, T.; Zhu, F.; Skrynnikov, N. R. Role of Electrostatic Interactions in Binding of Peptides and Intrinsically Disordered Proteins to Their Folded Targets. 2. The Model of Encounter Complex Involving the Double Mutant of the c-Crk N-SH3 Domain and Peptide Sos. *Biochemistry* **2016**, *55*, 1784–1800.

(96) Takemura, K.; Kitao, A. Water Model Tuning for Improved Reproduction of Rotational Diffusion and NMR Spectral Density. *J. Phys. Chem. B* **2012**, *116*, 6279–6287.

(97) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

(98) Karamertzanis, P. G.; Raiteri, P.; Galindo, A. The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration. *J. Chem. Theory Comput.* **2010**, *6*, 1590–1607.

(99) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(100) Hu, H.; Lu, Z.; Yang, W. Fitting Molecular Electrostatic Potentials from Quantum Mechanical Calculations. *J. Chem. Theory Comput.* **2007**, *3* (3), 1004–1013.

(101) Jarymowycz, V. A.; Stone, M. J. Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences. *Chem. Rev.* **2006**, *106*, 1624–1671.

(102) Piana-Agostinetti, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

(103) Vanommeslaeghe, K.; Yang, M.; MacKerell Jr., A. D. Robustness in the Fitting of Molecular Mechanics Parameters. *J. Comput. Chem.* **2015**, *36*, 1083–1101.

(104) Hopkins, C. W.; Roitberg, A. E. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *J. Chem. Inf. Model.* **2014**, *54* (7), 1978–1986.

(105) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.

(106) Ho, B. K.; Brasseur, R. The Ramachandran Plots of Glycine and Pre-Proline. *BMC Struct. Biol.* **2005**, *5*, 14.

(107) Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell Jr., A. D. Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, *9* (12), 5430–5449.

(108) Shi, Y.; Xia, Z.; Zhang, J.; Best, R. B.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4046–4063.

(109) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J. C.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater Scalability for Parallel Molecular Dynamics. *J. Comput. Phys.* **1999**, *151*, 283–312.

(110) Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell Jr., A. D.; Schulten, K.; Roux, B. High-Performance Scalable Molecular Dynamics Simulations of a Polarizable Force Field Based on Classical Drude Oscillators in NAMD. *J. Phys. Chem. Lett.* **2011**, *2*, 87–92.

(111) Wu, J. C.; Chattree, G.; Ren, P. Automation of AMOEBA Polarizable Force Field

Parameterization for Small Molecules. *Theor. Chem. Acc.* **2012**, *131*, 1138.

(112) Dunning Jr., T. H.; Peterson, K. A.; Wilson, A. K. Gaussian Basis Sets for Use in Correlated Molecular Calculations. X. The Atoms Aluminum through Argon Revisited. *J. Chem. Phys.* **2001**, *114* (21), 9244–9253.

(113) Woon, D. E.; Dunning Jr., T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum through Argon. *J. Chem. Phys.* **1993**, *98* (2), 1358–1371.

(114) Kendall, R. A.; Dunning Jr., T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96* (9), 6796–6806.

(115) Dunning Jr., T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023.

(116) Neese, F.; Wennmohs, F.; Becker, U.; Bykov, D.; Ganyushin, D.; Hansen, A.; Izsák, R.; Liakos, D. G.; Kollmar, C.; Kossmann, S.; Pantazis, D. A.; Petrenko, T.; Reimann, C.; Riplinger, C.; Roemelt, M.; Sandhöfer, B.; Schapiro, I.; Sivalingam, K.; Wezisla, B.; Kállay, M.; Grimme, S.; Valeev, E.; Chan, G. Orca 3.0.3. 2015.

(117) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383–394.

(118) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16* (11), 1339–1350.

(119) Grossfield, A. WHAM: The Weighted Histogram Analysis Method. Version 2.0.9. http://membrane.urmc.rochester.edu/content/wham.

(120) Song, K.; Stewart, J. M.; Fesinmeyer, R. M.; Andersen, N. H.; Simmerling, C. Structural Insights for Designed Alanine-Rich Helices: Comparing NMR Helicity Measures and Conformational Ensembles from Molecular Dynamics Simulation. *Biopolymers* **2008**, *89* (9), 747–760.

(121) Shalongo, W.; Dugad, L.; Stellwagen, E. Distribution of Helicity within the Model Peptide Acetyl(AAQAA)3amide. *J. Am. Chem. Soc.* **1994**, *116* (5), 8288–8293.

(122) Blanco, F. J.; Rivas, G.; Serrano, L. A Short Linear Peptide That Folds into a Native Stable β-Hairpin in Aqueous Solution. *Nat. Struct. Mol. Biol.* **1994**, *1* (9), 584–590.

(123) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 Residue Folded Peptide Designed by Segment Statistics. *Structure* **2004**, *12* (8), 1507–1518.

(124) Honda, S.; Akiba, T.; Kato, Y. S.; Sawada, Y.; Sekijima, M.; Ishimura, M.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K. Crystal Structure of a Ten-Amino Acid Protein. *J. Am. Chem. Soc.* **2008**, *130* (46), 15327–15331.

(125) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-Residue Protein. *Nat. Struct. Biol.* **2002**, *9* (6), 425–430.

(126) Reibarkh, M. Y.; Nolde, D. E.; I, V. L.; Bocharov, A. A.; Shulga, A. A.; Kirpichnikov, M. P.; Arseniev, A. S. Three-Dimensional Structure of Binase in Solution. *FEBS Lett.* **1998**, *431*, 250–254.

(127) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. Structure of Bovine Pancreatic Trypsin Inhibitor. Results of Joint Neutron and X-Ray Refinement of Crystal Form II. *J. Mol. Biol.* **1984**, *180*, 301–329.

(128) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. Evaluation of Backbone Proton Positions and Dynamics in a Small Protein by Liquid Crystal NMR Spectroscopy. *J. Am. Chem. Soc.* **2003**, *125* (30), 9179–9191.

(129) Walsh, M. A.; Schneider, T. R.; Sieker, L. C.; Dauter, Z.; Lamzin, S. Refinement of Triclinic Hen Egg-White Lysozyme at Atomic Resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1998**, *54*, 522–546.

(130) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of Ubiquitin Refined at 1.8 Å Resolution. *J. Mol. Biol.* **1987**, *194* (3), 531–544.

(131) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-Microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359* (3), 546–553.

(132) Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **1996**, *274*, 948–953.

(133) Kim, E. E.; Varadarajan, R.; Wyckoff, H. W.; Richards, F. M. Refinement of the Crystal Structure of Ribonuclease S. Comparison with and between the Various Ribonuclease A Structures. *Biochemistry* **1992**, *31* (49), 12304–12314.

(134) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with Amber - Part I: Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

(135) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(136) Miyamoto, S.; Kollman, P. A. Settle: An Analytical Version of the SHAKE and RATTLE

Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13* (8), 952–962.

(137) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

(138) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

(139) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

(140) Prompers, J. J.; Brüschweiler, R. General Framework for Studying the Dynamics of Folded and Nonfolded Proteins by NMR Relaxation Spectroscopy and MD Simulation. *J. Am. Chem. Soc.* **2002**, *124* (16), 4522–4534.

(141) Veenstra, D. L.; Ferguson, D. M.; Kollman, P. A. How Transferable Are Hydrogen Parameters in Molecular Mechanics Calculations? *J. Comput. Chem.* **1992**, *13* (8), 971–978.

(142) MacKerell Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Kiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *5647* (97), 3586–3616.

(143) Ting, D.; Wang, G.; Shapovalov, M. V; Mitra, R.; Jordan, M. I.; Dunbrack Jr., R. L. Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model. *PLoS Comput. Biol.* **2010**, *6* (4), e1000763.

(144) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer Jr., E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. *Eur. J. Biochem.* **1977**, *80* (2), 319–324.

(145) Best, R. B.; Buchete, N.-V.; Hummer, G. Are Current Molecular Dynamics Force Fields Too Helical? *Biophys. J.* **2008**, *95* (1), L07–L09.

(146) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *J. Am. Chem. Soc.* **2007**, *129* (5), 1179–1189.

(147) Hennig, M.; Bermel, W.; Schwalbe, H.; Griesinger, C. Determination of ψ Torsion Angle Restraints from 3J(Cα,Cα) and 3J(Cα,HN) Coupling Constants in Proteins. *J. Am. Chem. Soc.* **2000**, *122* (26), 6268–6277.

(148) Wirmer, J.; Schwalbe, H. Angular Dependence of 1J(Ni,Cαi) and 2J(Ni,Cα(i-1)) Coupling Constants Measured in J-Modulated HSQCs. *J. Biomol. NMR* **2002**, *23* (1), 47–55.

(149) Ding, K.; Gronenborn, A. M. Protein Backbone 1HN-13Cα and 15N-13Cα Residual Dipolar and J Couplings: New Constraints for NMR Structure Determination. *J. Am. Chem. Soc.* **2004**, *126*, 6232–6233.

(150) Case, D. A.; Scheurer, C.; Brüschweiler, R. Static and Dynamic Effects on Vicinal Scalar J Couplings in Proteins and Peptides: A MD/DFT Analysis. *J. Am. Chem. Soc.* **2000**, *122*, 10390–10397.

(151) Lindorff-Larsen, K.; Best, R. B.; Vendruscolo, M. Interpreting Dynamically-Averaged Scalar Couplings in Proteins. *J. Biomol. NMR* **2005**, *32* (4), 273–280.

(152) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. Folding Dynamics and Mechanism of β-Hairpin Formation. *Nature* **1997**, *390*, 196–199.

(153) Sanz, E.; Vega, C.; Abascal, J. L. F.; MacDowell, L. G. Phase Diagram of Water from Computer Simulation. *Phys. Rev. Lett.* **2004**, *92* (25), 1–4.

(154) Vega, C.; Sanz, E.; Abascal, J. L. F. The Melting Temperature of the Most Common Models of Water. *J. Chem. Phys.* **2005**, *122* (11), 1–9.

(155) Paschek, D.; Day, R.; García, A. E. Influence of Water-Protein Hydrogen Bonding on the Stability of Trp-Cage Miniprotein. A Comparison between the TIP3P and TIP4P-Ew Water Models. *Phys. Chem. Chem. Phys.* **2011**, *13* (44), 19840–19847.

(156) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. Rotational Diffusion Anisotropy of Human Ubiquitin from 15N NMR Relaxation. *J. Am. Chem. Soc.* **1995**, *117* (50), 12562–12566.

(157) Polyakov, K. M.; Lebedev, A. A.; Okorokov, A. L.; Panov, K. I.; Schulga, A. A.; Pavlovsky, A. G.; Karpeisky, M. Y.; Dodson, G. G. The Structure of Substrate-Free Microbial Ribonuclease Binase and of Its Complexes with 3′GMP and Sulfate Ions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58* (5), 744–750.

(158) Basconi, J. E.; Shirts, M. R. Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2887–2899.

(159) Gu, Y.; Li, D.-W.; Brüschweiler, R. NMR Order Parameter Determination from Long Molecular Dynamics Trajectories for Objective Comparison with Experiment. *J. Chem. Theory Comput.* **2014**, *10*, 2599–2607.

(160) Hall, J. B.; Fushman, D. Variability of the 15N Chemical Shielding Tensors in the B3 Domain of Protein G from 15N Relaxation Measurements at Several Fields. Implications for Backbone Order Parameters. *J. Am. Chem. Soc.* **2006**, *128* (24), 7855–7870.

(161) Lee, A. L.; Wand, A. J. Assessing Potential Bias in the Determination of Rotational Correlation Times of Proteins by NMR Relaxation. *J. Biomol. NMR* **1999**, *13* (2), 101–112.

(162) Zondlo, S. C.; Lee, A. E.; Zondlo, N. J. Determinants of Specificity of MDM2 for the Activation Domains of p53 and p65: Proline27 Disrupts the MDM2-Binding Motif of p53. *Biochemistry* **2006**, *45* (39), 11945–11957.

(163) Richards, F. M.; Vithayathil, P. J. The Preparation of Substilisin-Modified Ribonuclease and the Separation of the Peptide and Protein Components. *J. Biol. Chem.* **1959**, *234* (6), 1459–1465.

(164) Lovell, S. C.; Davis, I. W.; Adrendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure Validation by Cα Geometry: Φ, ψ and Cβ Deviation. *Proteins* **2003**, *50* (3), 437–450.

(165) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10* (11), 5113–5124.

(166) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(167) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. *Biopolymers* **2001**, *60* (2001), 96–123.

(168) Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J. Chem. Theory Comput.* **2014**, *10* (7), 2658–2667.

(169) Dickson, A.; Brooks, C. L. WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B* **2014**, *118* (13), 3532–3542.

(170) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* **1996**, *70* (1), 97–110.

(171) Jorgensen, W. L.; Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J. Chem. Theory Comput.* **2012**, *8* (10), 3895–3901.

(172) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2015**, *12*, 281–296.

(173) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38.

(174) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29* (11), 1859–1865.

(175) Brooks, B. R.; Brooks, C. L.; MacKerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaeffer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

(176) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621.

(177) Ponder, J. W. TINKER 7.1.2. 2015.

(178) Berendsen, H. J. C.; Postma, P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(179) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V; Cioslowski, J.; Fox, D. J. Gaussian 09. Gaussian, Inc.: Wallingform CT, 2009.

(180) Stone, A. J.; Alderton, M. Distributed Multipole Analysis. *Mol. Phys.* **1985**, *56* (5), 1047–1064.

(181) Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1* (6), 1128–1132.

(182) Lippert, R. A.; Predescu, C.; Ierardi, D. J.; Mackenzie, K. M.; Eastwood, M. P.; Dror, R.

O.; Shaw, D. E. Accurate and Efficient Integration for Molecular Dynamics Simulations at Constant Temperature and Pressure. *J. Chem. Phys.* **2013**, *139* (16), 164106.

(183) Pang, Y.; Buck, M.; Zuiderweg, E. R. P. Backbone Dynamics of the Ribonuclease Binase Active Site Area Using Multinuclear (15N and 13CO) NMR Relaxation and Computational Molecular Dynamics. *Biochemistry* **2002**, *41* (8), 2655–2666.

(184) Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **2016**, *12* (8), 3926–3947.

(185) Mittal, J.; Yoo, T. H.; Georgiou, G.; Truskett, T. M. Structural Ensemble of an Intrinsically Disordered Polypeptide. *J. Phys. Chem. B* **2013**, *117* (1), 118–124.

(186) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U.; Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinforma.* **2006**.

(187) Nguyen, H. T.; Pabit, S. A.; Meisburger, S. P.; Pollack, L.; Case, D. A.; Nguyen, H. T.; Pabit, S. A.; Meisburger, S. P.; Pollack, L. Accurate Small and Wide Angle X-Ray Scattering Profiles from Atomic Models of Proteins and Nucleic Acids Accurate Small and Wide Angle X-Ray Scattering Profiles from Atomic Models. *J. Chem. Phys.* **2014**, *141*, 1–15.

(188) Svergun, D. I.; Barberato, C.; Koch, M. H. J. CRYSOL - a Program to Evaluate X-Ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28* (6), 768–773.

(189) Petoukhov, M. V; Franke, D.; Shkumatov, A. V; Tria, G.; Kikhney, A. G.; Gajda, M.; Gorba, C.; Mertens, H. D. T.; Konarev, P. V; Svergun, D. I. New Developments in the ATSAS Program Package for Small-Angle Scattering Data Analysis. *J. Appl. Crystallogr.* **2012**, *45* (2), 342–350.

(190) Konarev, P. V; Volkov, V. V; Sokolova, A. V; Koch, M. H. J.; Svergun, D. I. PRIMUS: A Windows PC-Based System for Small-Angle Scattering Data Analysis. *J. Appl. Crystallogr.* **2003**, *36*, 1277–1282.

(191) Bax, A.; Grzesiek, S. Methodological Advances in Protein NMR. *Acc. Chem. Res.* **1993**, *26* (4), 131–138.

(192) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, M.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins Struct. Funct. Genet.* **2005**, *59* (4), 687–696.

(193) Palmer III, A. G. NMR Probes of Molecular Dynamics: Overview and Comparison with Other Techniques. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 129–155.

(194) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* **1995**, *6* (3), 277–293.

(195) D'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models I. Minimisation Algorithms and Their Performance within the Model-Free and Brownian Rotational Diffusion Spaces. *J. Biomol. NMR* **2008**, *40* (2), 107–119.

(196) D'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models II. A New Methodology for the Dual Optimisation of the Model-Free Parameters and the Brownian Rotational Diffusion Tensor. *J. Biomol. NMR* **2008**, *40* (2), 121–133.

(197) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Marius Clore, G. The Xplor-NIH NMR Molecular Structure Determination Package. *J. Magn. Reson.* **2003**, *160* (1), 65–73.

(198) Schwieters, C. D.; Clore, G. M. Using Small Angle Solution Scattering Data in Xplor-NIH Structure Calculations. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *80*, 1–11.

(199) Iwahara, J.; Schwieters, C. D.; Clore, G. M. Ensemble Approach for NMR Structure Refinement against 1H Paramagnetic Relaxation Enhancement Data Arising from a Flexible Paramagnetic Group Attached to a Macromolecule. *J. Am. Chem. Soc.* **2004**, *126*, 5879–5896.

(200) Clore, G. M.; Iwahara, J. Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. *Chem. Rev.* **2009**, *109*, 4108–4139.

(201) Schwieters, C. D.; Kuszewski, J. J.; Marius Clore, G. Using Xplor-NIH for NMR Molecular Structure Determination. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48* (1), 47–62.

(202) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. Ben; Siddique, N.; Snyder, C. W.; Spengler, J.; Tak, P.; Tang, P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2 : Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage, and Analysis*; 2014; pp 41–53.

(203) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J. C.; Tajkhorshid, E.; Villa, E.; Chipot, C.;

Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.

(204) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; Mackerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14* (1), 71–73.

(205) Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. FoXS, FoXSDock and MultiFoXS: Single-State and Multi-State Structural Modeling of Proteins and Their Complexes Based on SAXS Profiles. *Nucleic Acids Res.* **2016**, No. Figure 1, 1–6.

(206) Helmus, J. J.; Jaroniec, C. P. Nmrglue: An Open Source Python Package for the Analysis of Multidimensional NMR Data. *J. Biomol. NMR* **2013**, *55* (4), 355–367.

(207) McKinney, W. Pandas: A Foundational Python Library for Data Analysis and Statistics. *Python High Perform. Sci. Comput.* **2011**.

(208) Folk, M.; Heber, G.; Koziol, Q.; Pourmal, E.; Robinson, D. An Overview of the HDF5 Technology Suite and Its Applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*; 2011.

(209) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 99–104.

(210) Ben-Kiki, O.; Evans, C. YAML Ain't Markup Language Version 1.1. 2009.

(211) Debiec, K. T.; Chong, L. T.; Gronenborn, A. M. MolDynPlot. 2017.

(212) Berkholz, D. S.; Shapovalov, M. V; Dunbrack Jr., R. L.; Karplus, P. A. Conformation Dependence of Backbone Geometry in Proteins. *Structure* **2009**, *17* (10), 1316–1325.

(213) Berkholz, D. S.; Driggers, C. M.; Shapovalov, M. V; Dunbrack Jr., R. L.; Karplus, P. A. Nonplanar Peptide Bonds in Proteins Are Common and Conserved but Not Biased toward Active Sites. *Proc. Natl. Acad. Sci.* **2012**, *109* (2), 449–453.

(214) Debiec, K. T.; Chong, L. T.; Gronenborn, A. M. Ramaplot. pp 2–5.

(215) Alexandrescu, A. T.; Rathgeb-Szabo, K.; Rumpel, K.; Jahnke, W.; Schulthess, T.; Kammerer, R. A. 15N Backbone Dynamics of the S-Peptide from Ribonuclease A in Its Free and S-Protein Bound Forms: Toward a Site-Specific Analysis of Entropy Changes upon Folding. *Protein Sci.* **1998**, *7*, 389–402.

(216) Jaravine, V. A.; Alexandrescu, A. T.; Grzesiek, S. Observation of the Closing of Individual Hydrogen Bonds during TFE-Induced Helix Formation in a Peptide. *Protein Sci.* **2001**, *10*, 943–950.

(217) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeerschd, T.; Zurek, E.; Hutchison, G.

R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminform.* **2012**, *4* (8), 1–17.

(218) Wang, T.; Cai, S.; Zuiderweg, E. R. P. Temperature Dependence of Anisotropic Protein Backbone Dynamics. *J. Am. Chem. Soc.* **2003**, *125* (28), 8639–8643.